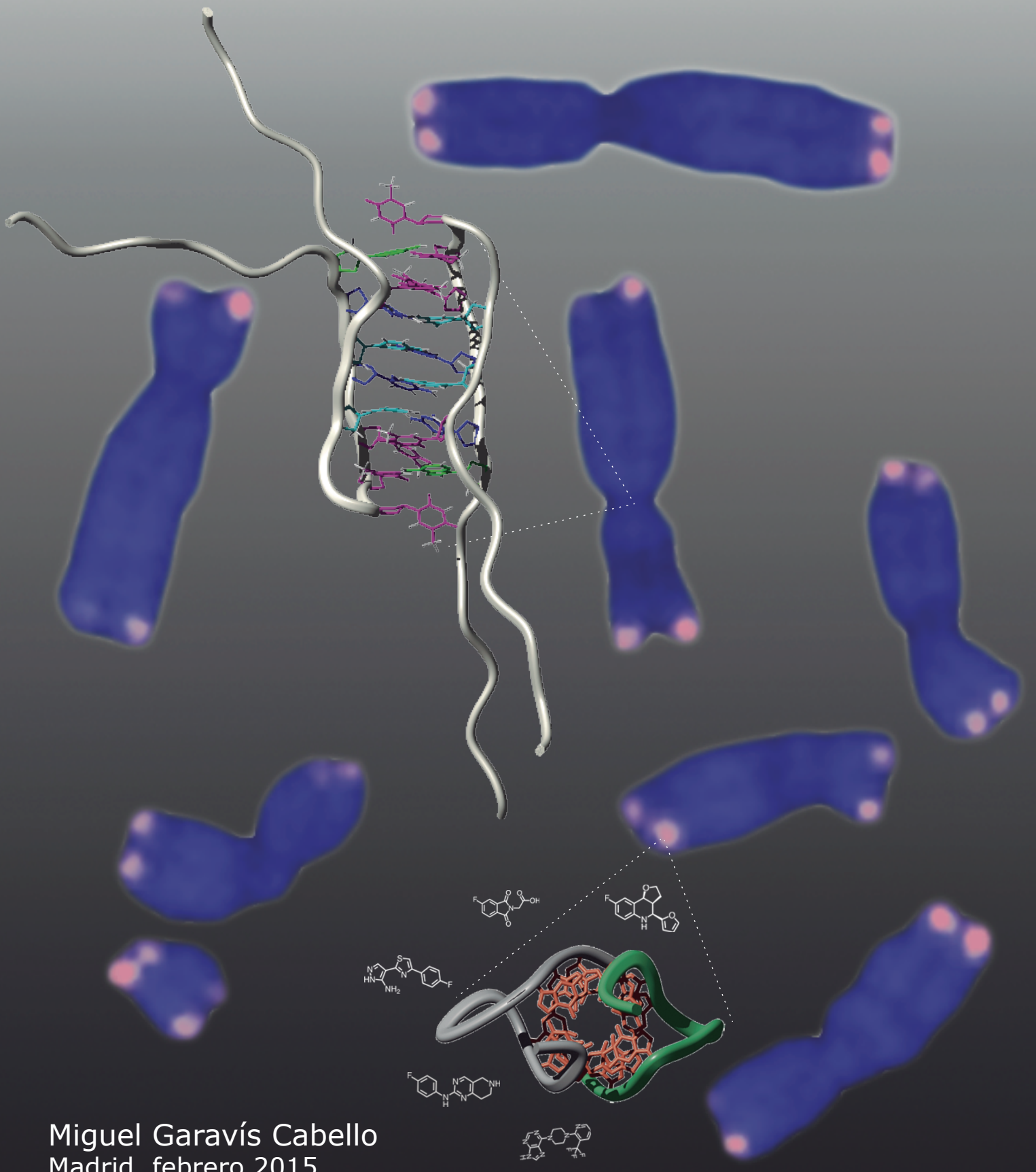
Universidad Autónoma de Madrid
Departamento de Biología Molecular
Facultad de Ciencias

Estructuras no canónicas de ácidos nucleicos en telómeros y centrómeros



Miguel Garavís Cabello
Madrid, febrero 2015

Departamento de Biología Molecular

Facultad de Ciencias

UNIVERSIDAD AUTÓNOMA DE MADRID

ESTRUCTURAS NO CANÓNICAS DE ÁCIDOS NUCLEICOS EN TELÓMEROS Y CENTRÓMEROS

Miguel Garavís Cabello

Licenciado en Química

Directores de Tesis

Dr. Alfredo Villasante Atienza

Dr. Carlos González Ibáñez

Tutor

Dr. Esteban Montejo de Garcini

Instituto de Química Física Rocasolano (CSIC)
Centro de Biología Molecular "Severo Ochoa" (UAM-CSIC)

RESUMEN

Los telómeros y los centrómeros son regiones del cromosoma que se caracterizan por la naturaleza repetitiva de las secuencias de DNA que las forman. En esta tesis se presenta una hipótesis sobre el origen y evolución de los telómeros y centrómeros. La hipótesis defiende que los primeros telómeros surgieron como consecuencia de la proliferación de retrotransposones en tándem en los finales de los primeros cromosomas lineales. El contenido en residuos de guanina de estos elementos móviles permitiría eventualmente la protección del final del cromosoma mediante la formación de estructuras no canónicas de DNA. Estos telómeros originales deberían poder cumplir una función doble que implica, tanto la protección del final del cromosoma, como la segregación del mismo durante la división celular. Más tarde, los centrómeros surgirían a partir de secuencias de las regiones subteloméricas. Estudios estructurales de secuencias centroméricas y teloméricas, como los abordados en esta tesis, son clave para la verificación de esta hipótesis.

El RNA telomérico (o TERRA) es un RNA no codificante que forma parte de la heterocromatina telomérica. Los datos acerca de su estructura son escasos y se limitan a moléculas formadas por un número reducido de repeticiones. En este trabajo se ha llevado a cabo el estudio estructural de secuencias de TERRA compuestas por un número de repeticiones teloméricas cercano al que se encuentra en las moléculas de RNA telomérico presentes en la célula. La caracterización estructural de estas moléculas se ha realizado utilizando técnicas biofísicas como la RMN y el dicroísmo circular, y técnicas de molécula individual como las pinzas ópticas. Así ha sido posible determinar que las moléculas de TERRA estudiadas se pliegan formando G-quadruplexes consecutivos que interaccionan entre sí. Por otro lado, debido al interés de TERRA como diana terapéutica, se han utilizado oligonucleótidos de TERRA para su aplicación en la búsqueda de pequeños compuestos fluorados capaces de unir RNA telomérico. En esta tesis se presenta la metodología, basada en ^{19}F-RMN, utilizada para la búsqueda de compuestos que unen TERRA así como las diferentes estrategias usadas para evaluar la afinidad y selectividad de los ligandos encontrados.

Las secuencias de DNA centroméricas son el resultado de la repetición de secuencias simples o complejas denominadas DNA satélite. En esta tesis se ha abordado el estudio estructural de secuencias pertenecientes a satélites complejos, como es el caso del satélite alfoide humano, y de secuencias derivadas de satélites simples, como el satélite dodeca de *Drosophila melanogaster*. En el presente trabajo se muestra la estructura de alta resolución de una versión truncada de la secuencia A-box perteneciente al satélite alfoide humano. Además se muestran datos estructurales que indican que las versiones no truncadas adoptan también estructuras diméricas tipo i-motif a pH ácido. Por último, en esta tesis se presenta la estructura del centrómero del cromosoma 3 de *D. melanogaster*. El análisis estructural de secuencias derivadas de la hebra rica en citosinas del satélite dodeca muestra que estas secuencias son también capaces de formar i-motifs. Estos resultados sugieren un posible papel de la estructura i-motif en la formación de la cromatina centromérica y plantean una posible explicación a la paradoja del centrómero.

Esta memoria se ha realizado de acuerdo a las normas de extensión y formato vigentes en el departamento de Biología Molecular de la Universidad Autónoma de Madrid en diciembre de 2014.

SUMMARY

Telomeres and centromeres are chromosomal regions characterised by the repetitive nature of their DNA sequences. In this thesis, we present a hypothesis on the origin and evolution of telomeres and centromeres. Our hypothesis argues that the first telomeres emerged as a consequence of the proliferation of tandemly arranged retrotransposons at the end of the nascent linear chromosomes. The guanine residue content of these mobile elements would eventually lead to protection or "capping" by means of formation of non-canonical DNA structures. These primal telomeres should be able to fulfil a dual function of protecting the end of the chromosome and allowing the correct segregation of the DNA during cell division. Later on, proper centromeres evolved at subtelomeric regions. Structural studies of telomeric and centromeric sequences, as those addressed in this thesis, are key elements to confirm the veracity of this hypothesis.

Telomeric RNA or TERRRA is a non-coding RNA that acts as a scaffold for the formation of telomeric heterochromatin. The information about the structure of TERRA is scarce and limited to molecules with a small number of repeats. In this work, we have carried out the structural study of TERRA sequences composed by a number of telomeric repeats that approaches the size of the nuclear endogenous telomeric RNA. The structural characterization of these molecules has been performed by using biophysical techniques as NMR and circular dichroism, and single molecule techniques as the optical tweezers. Thus, it has been possible to determine that the studied TERRA molecules fold into consecutive G-quadruplexes that interact with each other. Furthermore, given the importance of these molecules as therapeutic targets, we have used TERRA oligonucleotides for the screening of small fluorinated fragments that interact with telomeric RNA. Here, we present a ^{19}F-NMR based methodology for the screening of compounds that bind TERRA. In addition, we show the different strategies employed to analyse the affinity and selectivity of the hits identified in the screening.

Centromeric DNA is the result of the recurrence of simple or complex repeated sequences known as satellite DNA. In this thesis, we have carried out the structural study of both sequences belonging to complex satellites, as the human alfa-satellite, and sequences from simple centromeric satellites, as the dodeca satellite of *Drosophila melanogaster*. In the present work, we present the high-resolution structure of a truncated version of the A-box sequence, a sequence that appears in the human alfa-satellite. Besides, we show structural evidences indicating that complete versions of the A-box also fold into dimeric i-motif structures at acidic pH. Finally, we show the structure of the centromere of the chromosome 3 of *D. melanogaster*. The structural analysis of sequences derived from the C-rich strand of dodeca satellite shows that these sequences are also able to form i-motif structures. These results suggest a potential role of the i-motif in the formation of centromeric chromatin and also pose a possible explanation of the centromere paradox.

ÍNDICE

CLAVE DE ABREVIATURAS

ALT: *Alternative Lengthening of Telomeres*

BAC: *Bacterial Artificial Chromosome*

CENP: *Centromeric Protein*

COSY: *COrrelated Spectroscopy*

CPMG: Car-Purcel-Meiboom-Gill

CSA: *Chemical Shift Anisotropy*

d6-DMSO: Dimetilsulfoxido deuterado

DAPI: *4'. 6-diamino-2-phenilindole*

DC/CD: Dicroísmo Circular/*Circular Dichroism*

DCl: *Deuterium chloride*

DQF-COSY: *Double-Quantum-Filtered COSY*

DSBs: *Double Strand Breaks*

DSS: *2,2-dimetil-2-silapentano-5-sulfonato sódico*

EDTA: Ácido etilendiaminetetraacético

ESI: *Electrospray Ionization*

ESI-MS: *Electrospray Ionization Mass Spectrometry*

FBDD: *Fragment-Based Drug Discovery*

FISH: *Fluorescence In Situ Hybridization*

FPLC: *Fast Performance Liquid Chromatography*

FRET: *Fluorescence Resonance Energy Transfer*

HJURP: *Holliday Junction Recongnition Protein*

IGS: *Intergenic Spacer*

ILPR: *Insuline-Linked Polymorphic Region*

K_D: Constante de disociación

LE: *Ligand Efficiency*

LEF: *Local Environment of Fluorine*

MS: *Mass Spectrometry*

NOE: *Nuclear Overhauser Effect*

NOESY: *Nuclear Overhauser Effect Spectroscopy*

Non-LTR: *non-Long Terminal Repeat*

NOR: *Nucleolar Organizer Region*

OT: *Optical Tweezers*

PBS: *Phosphate Buffered Saline buffer*

PBST: *Phosphate Buffered Saline buffer with detergent Tween-20*

PDB: *Protein Data Bank*

PFA: *Paraformaldehyde*

PFGE: *Pulsed-Field Gel Electroforesis*

POT1: *Protection Of Telomeres 1*

ppm: partes por millón

PSA: *Polar Surface Area*

rDNA: *Ribosomal* DNA

RMN/NMR: Resonancia Magnética Nuclear/*Nuclear Magnetic Resonance*

RMSD: *Root Mean Square Deviation*

RPA: *Replication Protein A*

RT: *Reverse Transcriptase*

SPAM: *Solubility Purity and Aggregation of the Molecule*

SSC: *Saline Sodium Citrate buffer*

STD: *Saturation Transfer Difference*

TAS: *Telomere Asociated Sequences*

TERRA: *TElomeric Repeat-containing RNA*

TFA: *Trifluoroacetic Acid*

TOCSY: *TOtal Correlation Spectroscopy*

TRF1: *Telomeric Repeat-binding Factor 1*

TRF2: *Telomeric Repeat-binding Factor 2*

tRNA: *Transfer* RNA

Introducción

INTRODUCCIÓN

1 Estructura de ácidos nucleicos.

1.1 Estructura primaria de ácidos nucleicos

Los ácidos nucleicos, DNA y RNA, son polímeros cuya unidad de repetición se denomina nucleótido. Los nucleótidos están compuestos por tres fragmentos moleculares: un azúcar, una base heterocíclica y un grupo fosfato (Figura 1a). Existen dos azúcares que pueden formar parte de los ácidos nucleicos; la β-D-ribosa es el azúcar presente en el RNA y la β-D-2'-desoxirribosa en el DNA (Figura 1b). El heterociclo es una base nitrogenada plana y aromática. Está unida mediante un enlace covalente a la posición $C_{1'}$ del azúcar. Las bases nitrogenadas se clasifican en dos grupos en función de su estructura: púricas (derivadas de un anillo de purina) y pirimidínicas (derivadas de un anillo de pirimidina). Las bases púricas son adenina (A) y guanina (G), y las pirimidínicas son citosina (C), timina (T) y uracilo (U) (Figura 1c). La base uracilo solo está presente en el RNA mientras que la base timina solo se encuentra en el DNA.

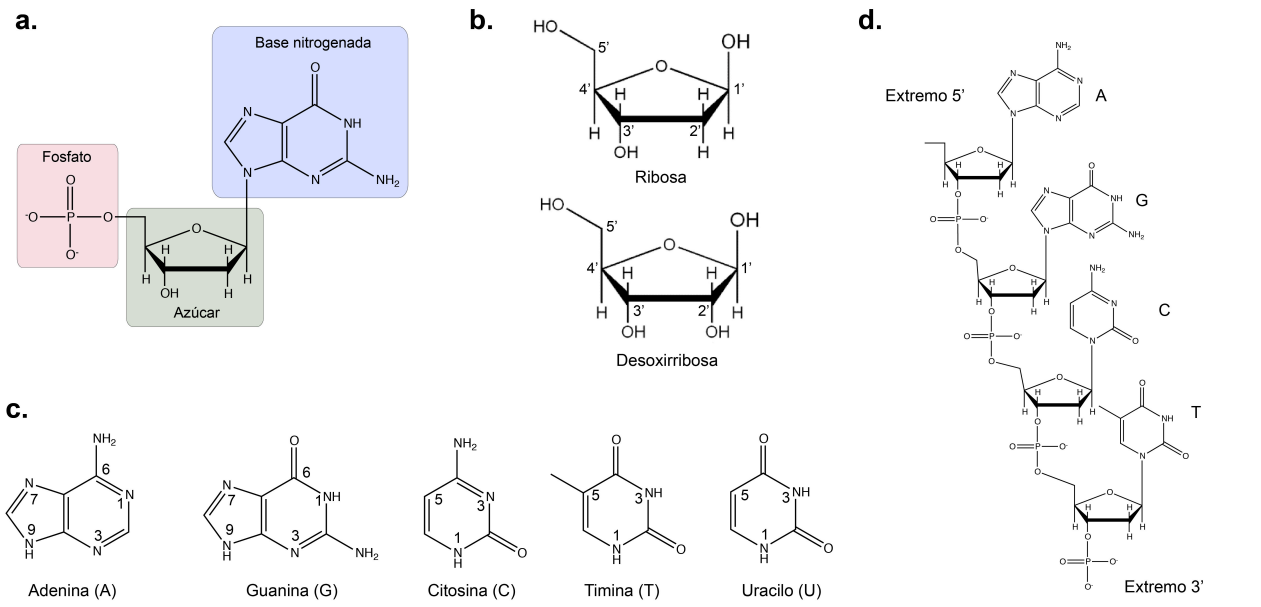


Figura 1. *(a) Estructura química de un nucleótido mostrando los componentes que lo integran: un azúcar (desoxirribosa) en verde, una base nitrogenada (guanina) en azul y un grupo fosfato en rojo. (b) Estructura química de los azúcares ribosa y desoxirribosa. (c) Estructura química de las bases nitrogenadas que forman parte de los ácidos nucleicos. (d) Cadena de nucleótidos en la que se muestran los enlaces fosfodiéster y los extremos de la molécula, denominados 5' y 3'.*

Los nucleótidos presentan un grupo fosfato en la posición $C_{5'}$ (Figura 1a). Este grupo fosfato está implicado en la formación del enlace fosfodiéster que une un nucleótido con el grupo hidroxilo 3' del nucleótido siguiente (Figura 1d). La carga negativa del enlace fosfodiéster es la causa de que los ácidos nucleicos sean polianiones, lo cual repercute de forma decisiva tanto en la estructura secundaria que adoptan como en su interacción con el medio que los rodea. Por lo tanto, los ácidos nucleicos son cadenas de nucleótidos con una parte repetitiva constituida por el esqueleto azúcar-fosfato y otra parte variable definida por la secuencia de bases nitrogenadas. Esta parte variable es la que contiene la información genética.

1.1.1. Parámetros conformacionales de nucleótidos

1.1.1.1 Conformación del anillo del azúcar

Una de las características que contribuye a la flexibilidad estructural de los ácidos nucleicos se debe a las diferentes conformaciones que puede adoptar el anillo de ribosa y desoxirribosa. Los cinco átomos del anillo del azúcar adoptan normalmente una disposición fuera del plano. Esta conformación del azúcar se denomina *puckering* y se puede clasificar en dos tipos: "silla" o *twist* (T) y "sobre" o *envelope* (E) (Blackburn, 2006).

Para describir de forma analítica las diferentes conformaciones del azúcar se define el concepto de pseudorrotación a través de dos parámetros: el ángulo de fase de pseudorrotación (P), que indica el tipo de conformación del anillo, y la amplitud (ν_m) que indica la magnitud máxima del mismo (Altona and Sundaralingam, 1973; Altona and Sundaralingam, 1972).

El ángulo de pseudorrotación (P) (Altona and Sundaralingam, 1972) a priori puede tomar valores que varían entre 0° y 360° (Figura 2a). Sin embargo, se ha comprobado experimentalmente que los anillos de desoxirribosa adoptan preferiblemente dos conformaciones: la C3'-endo o norte (N) y la C2'-endo o sur (S) (Figura 2b). La conformación norte tiene ángulos de pseudorrotación entre 1° y 40° mientras que la conformación sur presenta valores entre 130° y 180°.

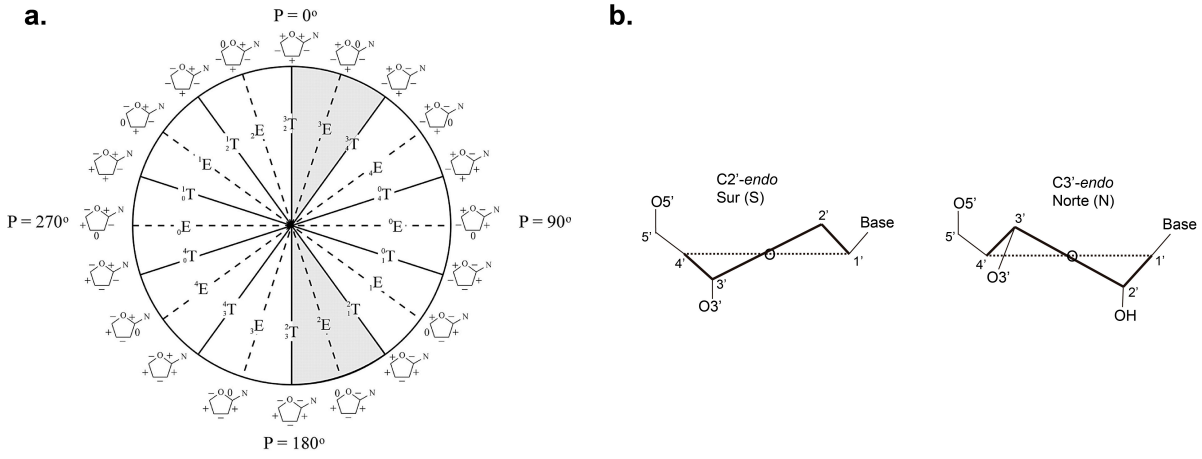


***Figura 2**. a. Ciclo de pseudorrotación del anillo de desoxirribosa, donde se muestra la relación entre la fase del ángulo de pseudorrotación (P) y las formas "sobre" (E) y "silla" (T). b. Conformaciones Norte (N) del anillo de ribosa (derecha) y Sur (S) (izquierda) del anillo de desoxirribosa.*

1.1.1.2 Conformación del ángulo glicosídico

El ángulo glicosídico (χ) es el ángulo diedro definido por los átomos O4'-C1'-N9-C4 en las bases púricas y O4'-C1'-N1-C2 en las bases pirimidínicas. Este ángulo puede tomar un amplio rango de valores, sin embargo, las bases nitrogenadas adoptan dos orientaciones preferidas con respecto a la posición del azúcar: las denominadas *syn* y *anti* (Donohue and Trueblood, 1960; Haschemeyer and Rich, 1967) (Figura 3a).

En la conformación *anti* el azúcar está en posición opuesta a la base mientras que en la *syn*, el azúcar se encuentra eclipsando a la base. La conformación *anti* es por norma general la más estable en los ácidos nucleicos, aunque la conformación *syn* se observa con frecuencia en determinados motivos estructurales.

1.1.1.3 Conformación del esqueleto azúcar fosfato

Además de la conformación del azúcar y del ángulo glicosídico, la conformación global de una cadena polinucleotídica depende de los seis ángulos de torsión (α, β, γ, δ, ε y ζ) del esqueleto azúcar-fosfato (Figura 3b). Estos ángulos, no son independientes entre sí, por lo que cambios conformacionales en el polinucleótido, conllevan cambios concertados en varios de estos ángulos de torsión.

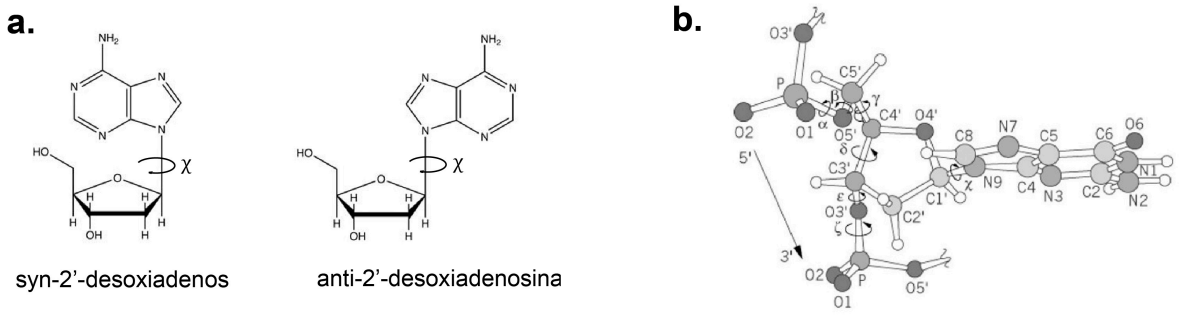


Figura 3. *(a) Conformaciones del ángulo glicosídico (syn y anti) en nucleósidos de adenosina. (b) Ángulos de torsión del esqueleto azúcar-fosfato de un nucleótido.*

1.1.2. Parámetros conformacionales helicoidales

Para describir la estructura de los ácidos nucleicos se pueden utilizar parámetros relacionados con la posición y geometría de los pares de bases con respecto al sistema de coordenadas definido por el eje de la hélice que forma el ácido nucleico objeto de estudio. Como se detallará más adelante, las cadenas de nucleótidos tienden a asociarse unas con otras a través de la formación de pares de bases. Son necesarios cuatro parámetros para definir la posición de un par de bases con respecto al eje z de la hélice (*y-displacement (dy)*, *x-displacement (dx)*, *inclination (η)* y *tip (θ)*), seis parámetros para definir la geometría de un par de bases (*stagger*, *strech*, *shear, opening*, *propeller twist* y *buckle)* y otros seis para definir la posición relativa de un par de bases con respecto al siguiente *(slide*, *shift*, *rise*, *twist*, *roll* y *tilt)*.

1.2. Estructura secundaria de ácidos nucleicos

1.2.1. Interacciones estabilizantes de estructura secundaria: enlaces de hidrógeno e interacciones de apilamiento

Las características físico-químicas de los nucleótidos hacen posible que estos interaccionen entre sí, posibilitando la formación de diferentes estructuras secundarias. Dos de las interacciones más importantes entre las bases son (i) las debidas a la formación de enlaces de hidrógeno y (ii) las debidas a interacciones de apilamiento entre las bases, que se deben a fuerzas de dispersión de London y a efectos hidrofóbicos.

Para la formación de un enlace de hidrógeno es necesario que un átomo de hidrógeno (dador) se una a dos átomos X e Y de mayor electronegatividad (aceptor). Los enlaces de hidrógeno que se establecen entre bases son del tipo N-H···N y N-H···O, pudiendo ser el grupo dador N-H tanto amino como imino.

Bajo el criterio de que es necesaria la formación de un mínimo de dos enlaces de hidrógeno para producir un par de bases estable, existen muy diversos modos de asociación entre pares de nucleótidos. Los

pares de bases se pueden dar entre nucleótidos con la misma base nitrogenada (*homo base-pairs*) o entre nucleótidos con bases diferentes (*hetero base-pairs*). Los pares de bases estándar, también llamados de Watson y Crick, están formados por una base púrica y otra pirimidínica: adenina con timina (o uracilo) y guanina con citosina. Los pares A·T se mantienen mediante dos enlaces de hidrógeno (Figura 4) mientras que en los pares G·C se forman tres. Esto hace que, en general, los pares G·C confieran mayor estabilidad que los pares A·T. Sin embargo, la distancia entre bases en los pares G·C y A·T es muy similar, lo que permite la formación de estructuras helicoidales denominadas canónicas (sección 1.2.2).

A medida que ha ido aumentando el número de estructuras de DNA determinadas experimentalmente, se han observado cada vez más estructuras que contienen pares de bases distintos a los de Watson y Crick. El conjunto de pares que se puede dar entre bases nitrogenadas se puede clasificar en Watson y Crick (WC), Watson y Crick reverso (rWC), Hoogsteen (H), Hoogsteen reverso (rH) (Figura 4), *Wobble (W)* y *missmatches*.

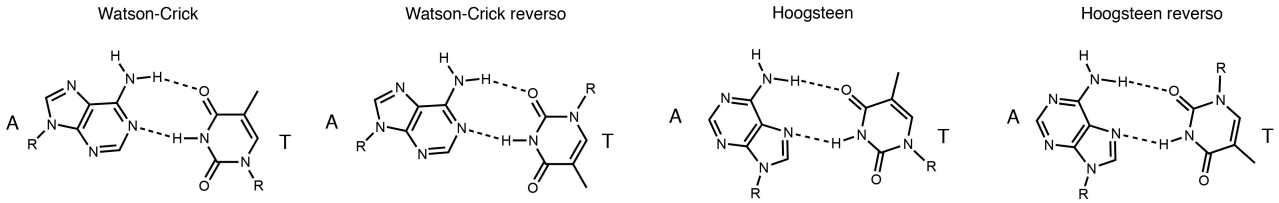


Figura 4. *Pares de bases A·T Watson y Crick reverso, A:U Wooble y A:T Hoogsteen y Hoogsteen reverso.*

Otro tipo de interacciones que estabilizan las estructuras secundarias de los ácidos nucleicos son las interacciones de apilamiento. Éstas, están relacionadas con el solapamiento de orbitales π de anillos aromáticos y con la interacción entre dipolos. Además, las interacciones hidrofóbicas y las fuerzas de dispersión de London también tienen una contribución muy importante en la estabilización de estructuras en disolución acuosa.

Pese a que se reconoce la existencia y notable influencia de las interacciones hidrofóbicas en la estructura de muchas biomoléculas, su origen aún es materia de debate. La hipótesis más apoyada indica que las moléculas no polares disueltas en un medio polar se apilan para minimizar su superficie expuesta a las moléculas del medio, lo que resulta en un aumento de entropía de las moléculas del disolvente (Franks, 1975).

1.2.2. Estructuras canónicas de ácidos nucleicos: Dúplex tipo A y tipo B.

Las estructuras que pueden adoptar el DNA y el RNA pueden ser clasificadas en dos grupos: las formas canónicas, que son las formadas por dos cadenas en orientación antiparalela que dan lugar a dobles hélices dextrógiras, y las formas no canónicas, que son aquellas que no cumplen la condición anterior. Dentro de las formas canónicas que puede adoptar una determinada secuencia de DNA podemos encontrar la forma A y la forma B (Figura 5), el RNA, por el contrario siempre forma hélice tipo A. Además, en

determinadas secuencias de nucleótidos y bajo ciertas condiciones experimentales se han encontrado otras estructuras, la mayoría de las cuales son subclases de las hélices tipo A y B.

La estructura más habitual del DNA en disolución es la doble hélice propuesta por Watson y Crick en 1953 a partir de datos de difracción de rayos X (Watson and Crick, 1953). Esta conformación del DNA también se conoce como B-DNA o DNA tipo B. Es el resultado de la asociación de dos cadenas complementarias y antiparalelas. Dicha asociación se produce a través de la formación de pares de bases entre bases complementarias (A·T y G·C). El tipo de apareamiento entre bases es el denominado de Watson y Crick (Figura 4), aunque recientemente se ha observado la formación transitoria de pares tipo Hoogsteen en dúplex canónicos de DNA (Nikolova et al., 2011). Los pares de bases se sitúan en el centro, en planos prácticamente perpendiculares al eje de la hélice quedando el esqueleto azúcar fosfato en la parte exterior. Esta disposición de los nucleótidos hace que en la estructura aparezcan dos surcos de distinta profundidad y anchura, que reciben los nombres de surco menor o *minor groove* y surco mayor o *mayor groove* (Figura 5).

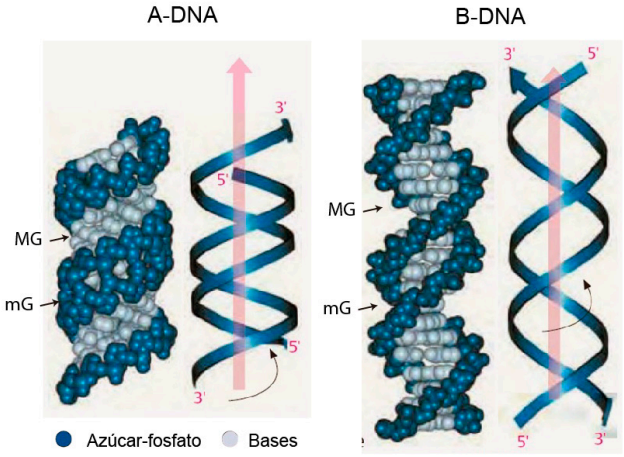


Figura 5. *Estructuras de las hélices tipo A y tipo B del DNA en las que se marcan la dirección de la hélice (flecha roja), el sentido de giro de la hélice (flecha curva negra) y los surcos mayor (MG) y menor (mG) de cada estructura.*

Las hélices de tipo A se asemejan a las de tipo B en que también son dextrógiras y presentan dos surcos de dimensiones diferentes. La doble hélice tipo A es más achatada y ancha que la tipo B. Además, los pares de bases presentan una inclinación con respecto al eje de la hélice (aprox. 20 °) lo que no ocurre en el DNA tipo B. Este tipo de estructura se observa en moléculas de DNA que han sido cristalizadas en condiciones de baja humedad y alta fuerza iónica así como en disolución en presencia de diversos alcoholes (Portmann et al., 1995). Por otro lado, la hélice de RNA siempre adopta la conformación tipo A, tanto cuando se asocian dos hebras de RNA como en híbridos DNA:RNA.

1.2.3. Estructuras no canónicas de los ácidos nucleicos

Como se ha dicho anteriormente, la forma más frecuente de encontrar el DNA en las células es como una hélice de dos hebras antiparalelas. Sin embargo, durante determinados proceso del metabolismo, tales como la replicación o la transcripción, el DNA de doble cadena se despliega parcialmente dando lugar a

regiones de una sola hebra. Algunas de estas regiones susceptibles de estar en forma de DNA de cadena sencilla muestran secuencias repetitivas. Bajo ciertas condiciones, estas secuencias de DNA repetitivo pueden formar estructuras diferentes a la doble hélice tipo B. Este tipo de plegamientos se denominan estructuras no canónicas. Algunas de las más relevantes son el Z-DNA, los dúplex tipo Hoogsteen (Abrescia et al., 2004), las horquillas (*hairpins*) de DNA, las hélices triples o tríplex, el DNA cruciforme, el dúplex de poliadeninas (A-motif) y los tétraplex. Estas estructuras no convencionales podrían estar involucradas en importantes procesos biológicos (Bacolla and Wells, 2009; Wang and Vasquez, 2006; Zhao et al., 2010). En esta introducción nos enfocaremos en aspectos referentes a las estructuras tipo tétraplex, sobre las cuales se han realizado los estudios descritos en esta tesis.

Tétraplex: cuádruplex de guaninas e *i-motif*

Los tétraplex engloban a aquellas estructuras en las que intervienen cuatro hebras de DNA o RNA pudiendo provenir dichas hebras de una misma molécula o de moléculas diferentes. Estas estructuras se pueden clasificar en dos grupos en función de qué bases se asocian entre sí para formar la estructura. Así se diferencian los cuádruplex de guanina, en los que hay apareamiento entre bases de guanina y los *i-motif*, en los que el apareamiento se produce entre bases de citosina.

1.2.3.1. Cuádruplex de guanina (*G-quadruplex*)

Los cuádruplex de guanina son estructuras de cuatro hebras de DNA o RNA formadas como consecuencia del apilamiento de al menos dos tétradas de guanina (Burge et al., 2006; Simonsson, 2001). Las tétradas de guanina son estructuras planas formadas por cuatro guaninas que se asocian entre sí mediante ocho enlaces de hidrógeno tipo Hoogsteen (Davis, 2004) (Figura 6a izquierda).

Los cuádruplex de guanina requieren de la presencia de cationes metálicos para su estabilización, especialmente metales alcalinos (y un orden de preferencia $K^+ > Na^+$). El efecto estabilizador de los cationes alcalinos se producen porque éstos establecen interacciones iónicas con los oxígenos en la posición 6 de las guaninas que forman las tétradas. Los iones K^+ y los iones amonio tienden a situarse entre dos tétradas consecutivas, coordinándose con los 8 átomos de oxígeno de las guaninas de las dos tétradas vecinas. Los cationes Na^+, cuyo radio atómico es menor pueden situarse en el centro de una tétrada coordinándose con los cuatro oxígenos O6 de una tétrada o situarse entre dos tétradas sucesivas (Horvath and Schultz, 2001; Schultze et al., 1999).

Los cuádruplex de guanina pueden presentar distinta molecularidad ya que pueden estar constituidos por una (cuádruplex monomérico o intramolecular), por dos (cuádruplex dimérico) (Figura 6a) o por cuatro (cuádruplex tetramolecular) moléculas de ácido nucleico.

Las secuencias que se encuentran entre los sucesivos tramos de guanina que forman las tétradas reciben el nombre de lazos o *loops* y permiten unir las tétradas apiladas de formas muy diversas, lo que provoca que haya una gran variedad de topologías para los cuádruplex de guanina. Estos *loops* pueden ser diagonales, laterales o de tipo *propeller* (o tipo aspa). Estos últimos conectan dos cadenas en la misma orientación, mientras que los diagonales y los laterales conectan cadenas con orientación opuesta. En función a la

orientación de las hebras se definen dos tipos de cuádruplex de guaninas: (i) cuádruplex paralelo, en el que todas las hebras se orientan en el mismo sentido (Figura 6a centro) y (ii) cuádruplex antiparalelo, en el que al menos una de las hebras se orienta en sentido contrario a las demás (Figura 6a derecha).

Los cuádruplex de guanina presentan cuatro surcos en su estructura, en contraposición a los dos que tienen las hélices dobles. Las dimensiones de los surcos son variables y dependientes de la topología y naturaleza de los *loops,* así como de los ángulos glicosídicos de los nucleótidos de guanina. Por otro lado, los cuádruplex de guanina pueden formar parte de estructuras de orden superior, (Chang et al., 2007; Lin et al., 2011; Smargiasso et al., 2008), donde la subunidad que se repite es un cuádruplex de guanina. Este tipo de estructuras parecen estar favorecidas por *loops* cortos apareciendo en estos casos incluso a bajas concentraciones de oligonucleótido.

1.2.3.2. I-motif

Los *i-motif* se forman por la asociación de dos dúplex paralelos mediante la formación de pares de bases de citosinas hemiprotonada $C:C^+$ que se intercalan de forma antiparalela (Figura 6b). El nombre *i-motif* (del inglés *intercalated motif,* y en español motivo intercalado) se le otorga por el hecho de que ésta es la única estructura conocida donde los pares de bases aparecen intercalados entre sí (Figura 6b derecha). El componente básico de esta estructura es el par de bases $C:C^+$ que está estabilizado por tres enlaces de hidrógeno (Figura 6b izquierda). En general, el hidrógeno que enlaza los nitrógenos N3 de las dos citosinas que forman el par de bases no está a la misma distancia de ambos nitrógenos, pero adopta una posición tal que se maximiza su distancia con respecto al hidrógeno del siguiente par $C:C^+$.

La disposición de los pares origina dos surcos anchos y planos y otros dos muy estrechos. De acuerdo al ordenamiento de los pares apilados se pueden distinguir dos clases de *i-motif*: aquellos cuyo par $C:C^+$ terminal está formado por citosinas del extremo 3' de la secuencia, llamados 3'E (Figura 6b centro) y aquellos en los que el par de bases terminal está formado por nucleótidos del extremo 5' llamados 5'E (Figura 6b derecha).

Al igual que ocurre con los cuádruplex de guanina, los *i-motif* pueden ser intramoleculares si se forman a partir de una sola secuencia que contenga los segmentos de citosina suficientes, o intermoleculares si se forman a partir de dos (Figura 6b) o cuatro moléculas de DNA donde cada una de ellas aporta las citosinas necesarias para formar la estructura.

La estabilidad de los *i-motif* depende de muchos factores tales como la secuencia de nucleótidos, la fuerza iónica o la temperatura, entre otros. Dado que la protonación de una de las citosinas es un requisito fundamental para la formación del *i-motif*, el pH tiene un papel crucial. El pK_a de la citosina es aproximadamente 4.6 (en agua pura y a 25 °C) por lo tanto se espera que la formación del *i-motif* ocurra a valores de pH menores de 6.6, aproximadamente (Mergny et al., 1995; Völker et al., 2007). Por tanto, en un rango de pH que abarca aproximadamente desde 4 hasta 7 y a 25 °C, las bases de citosina están parcialmente protonadas y el DNA se puede plegar formando un *i-motif*. A valores de pH más ácido (aproximadamente menores de 3) todas las citosinas están protonadas y no pueden formar el patrón de enlaces de hidrógeno necesario para la formación del par $C:C^+$. El *i-motif* no se forma en condiciones de

temperatura, pH y fuerza iónica semejantes a las fisiológicas, sin embargo, existen casos en los que su formación se ha detectado a bajas temperaturas y pH neutro (Escaja et al., 2012) o ligeramente básico (Zhou et al., 2010). Recientemente se ha demostrado la formación de estructuras *i-motif* en presencia de cationes de Ag^+ y pH fisiológico (Day et al., 2013). Existen además, ciertos procesos biológicos que pueden alterar localmente el pH de la célula. Por último, el ambiente celular debe tenerse también en cuenta. Algunos estudios muestran que en presencia de agentes de apiñamiento (o *crowding*) se pueden formar *i-motifs* a pH neutro (Rajendran et al., 2010).

En las estructuras *i-motif*, además de pares $C:C^+$ se pueden formar también apareamientos con otros nucleótidos que pueden acomodarse en la estructura. Así, por ejemplo las timinas pueden formar pares de bases simétricos T·T que son prácticamente isomorfos a los $C:C^+$ (Canalia and Leroy, 2009; Canalia and Leroy, 2005). Además en algunos casos, la interacción entre nucleótidos del *loop* da lugar a motivos estructurales como tétradas G:T:G:T que se han observado en algunas estructuras diméricas (Escaja et al., 2012; Gallego et al., 1997).

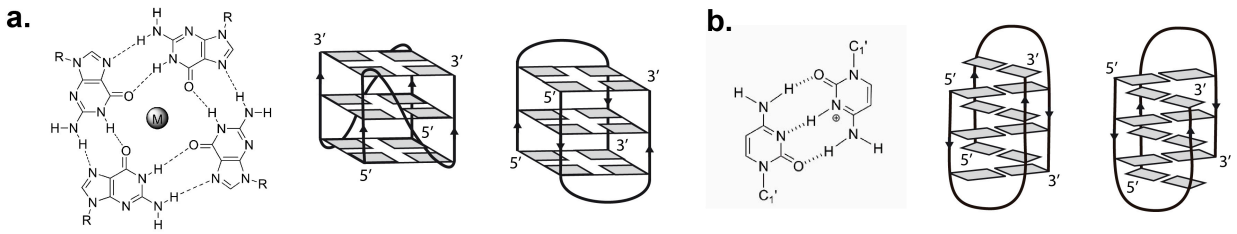


***Figura 6**. (a) Cuádruplex de guanina. Tétrada de guaninas (izquierda) y estructuras de cuádruplexes de guanina diméricos paralelo (centro) y antiparalelo (derecha). (b) I-motif. Par de bases $C:C^+$(izquierda) y estructuras de i-motifs diméricos con las conformaciones 3'E (centro) y 5'E (derecha).*

Del mismo modo que ocurre con los cuádruplex de guanina, los *i-motif* tienen la capacidad potencial de formar estructuras de orden superior que tienen gran interés en el campo de la nanotecnología. Así por ejemplo, existen estrategias para construir ensamblajes basados en el apilamiento de *i-motifs* y que reciben el nombre de *I-wires*.

2 Técnicas experimentales para el estudio de estructura de ácidos nucleicos.

2.1 Resonancia Magnética Nuclear

Las dos únicas técnicas experimentales que permiten obtener estructuras de ácidos nucleicos a nivel atómico son la cristalografía de rayos X y la Resonancia Magnética Nuclear (RMN). La cristalografía de rayos X se desarrolló con anterioridad y muchos de los primeros descubrimientos sobre la estructura de los ácidos nucleicos se deben a esta técnica. En la actualidad, es la única técnica capaz de determinar estructuras de alta resolución de ácidos nucleicos de tamaño relativamente grande (más de 100 nucleótidos). La RMN, en cambio, sólo permite la determinación de estructuras de secuencias de ácidos nucleicos de pequeño tamaño, pero tiene la ventaja de trabajar sobre muestras en disolución y, por tanto, en condiciones más parecidas a las fisiológicas. Aunque a lo largo de esta tesis se ha intentado obtener cristales de ácidos

nucleicos, la técnica que se ha empleado para la caracterización estructural ha sido la RMN.

La RMN se ha utilizado tradicionalmente para el análisis de la estructura de compuestos orgánicos. Sin embargo, avances técnicos como el desarrollo de imanes superconductores de alto campo, la introducción de la espectroscopia bi- y multidimensional o el progreso en el marcaje isotópico de macromoléculas, han hecho que desde mediados de la década de los 80, la RMN se utilice extensivamente como técnica de determinación estructural de macromoléculas biológicas. Mediante RMN se pueden determinar estructuras de macromoléculas en condiciones similares a las que se dan en la célula, y se puede extraer además, información dinámica de las estructuras estudiadas. Así, esta técnica permite estudiar fenómenos que tienen lugar en muy diversas escalas de tiempo, que van desde los picosegundos a los segundos. La información dinámica es especialmente importante en ácidos nucleicos, dado que suelen presentar regiones de alta flexibilidad conformacional. Estas regiones, a su vez, suelen estar involucradas en los procesos de reconocimiento molecular del DNA y RNA, para cuya caracterización la RMN es una herramienta de enorme utilidad.

La técnica de RMN se basa en el efecto del campo magnético sobre los núcleos atómicos que poseen espín, también denominados núcleos magnéticamente activos, ya que éstos se comportan como pequeños imanes. En ausencia de campo magnético, los espines se orientan al azar. Sin embargo, cuando una muestra se coloca en un campo magnético estático, los espines tienden a orientarse y aparece una magnetización macroscópica neta en la dirección del campo (magnetización longitudinal). Los núcleos con espín ½ pueden orientarse en la misma dirección o en dirección opuesta al campo magnético. Las dos direcciones se corresponden con estados de diferente energía, denominados α y β. En el equilibrio existen más núcleos en el estado α, de menor energía, que en el estado β. La particularidad de la RMN con respecto a otras técnicas espectroscópicas, consiste en que la diferencia de energía entre los dos estados es muy pequeña. Este efecto tiene dos consecuencias principales: (i) que son necesarias frecuencias de resonancia muy bajas, en el rango de las radiofrecuencias, y (ii) que la diferencia de poblaciones en los dos estados no puede ser muy grande. Esta pequeña variación de poblaciones en el estado α y β tiene como consecuencia que la técnica sea muy poco sensible en comparación con otras espectroscopías. En la práctica, este efecto provoca que sea necesario trabajar con muestras muy concentradas, típicamente en el rango milimolar. Por otra parte, como la diferencia de energía entre los dos estados de espín depende de la intensidad del campo magnético aplicado, cuanto mayor es el campo magnético, mayor diferencia de energía hay entre los dos estados de espín, y por tanto la diferencia de poblaciones entre ambos será también mayor. Este efecto tiene como consecuencia que la sensibilidad de la técnica aumente a campos magnéticos más intensos. Por ello, a lo largo de esta tesis hemos utilizado espectrómetros de alto campo (600, 700 y 800 MHz).

2.1.1 Información estructural obtenida en los experimentos de RMN

De los diversos experimentos de RMN se puede extraer diferente información experimental. Dicha información está asociada a los diferentes parámetros que se pueden obtener en los experimentos de RMN. Algunos de los más importantes para el estudio de estructuras de ácidos nucleicos son:

a) *Desplazamiento químico*:

El campo magnético neto experimentado por el núcleo no es exactamente el campo externo aplicado, ya que éste se ve modificado por el efecto de la nube electrónica que rodea al núcleo, que induce a su vez campos magnéticos que contrarrestan o refuerzan este campo externo. Este fenómeno es conocido como apantallamiento y depende del entorno químico en donde se encuentre cada núcleo. Se conoce con el nombre de desplazamiento químico (δ), al cambio o desplazamiento de la frecuencia de resonancia de un núcleo como consecuencia de la estructura química de la que forma parte. El desplazamiento químico se da en partes por millón (ppm) respecto a una referencia normalizada. De esta manera, en la escala de desplazamientos químicos la frecuencia de resonancia no depende del campo magnético externo y se pueden comparar espectros obtenidos en espectrómetros con diferentes campos.

El desplazamiento químico es el parámetro más sensible de los obtenidos por RMN, por esta razón, pequeñas variaciones de éste permiten determinar alteraciones en la estructura de la biomolécula estudiada. En el caso de ácidos nucleicos, las resonancias correspondientes a los protones imino de las bases nitrogenadas son particularmente útiles, puesto que aparecen hacia campos más bajos cuando se forma el enlace de hidrógeno. Estos desplazamientos químicos varían para cada tipo de apareamiento. Así, por ejemplo, en los pares no canónicos T·T las señales imino aparecen a campos mayores (10-11 ppm) mientras que, en el caso de pares de bases tipo Watson y Crick, lo hacen a campos más bajos (12-14 ppm). De este modo, los desplazamientos químicos son un claro indicador del tipo de estructura secundaria.

b) Constantes de acoplamiento (J):

Otra de las magnitudes observables mediante RMN es el llamado acoplamiento escalar. Se trata de una interacción magnética que se transmite por los electrones de enlace. Este fenómeno provoca que la frecuencia de resonancia de un núcleo se vea alterada por la presencia de otros núcleos magnéticos, que se encuentran directamente enlazados al primero o conectados con él a través de un número limitado de enlaces. Esta interacción da como resultado un desdoblamiento de las señales del espectro.

El acoplamiento escalar de espín se relaciona con un parámetro espectral denominado constante de acoplamiento escalar, que se denota con la letra J y que se extrae normalmente de los experimentos tipo COSY. Desde el punto de vista estructural, la constante de acoplamiento más importante es la constante a tres enlaces. El valor de esta constante depende del ángulo diedro definido por los tres enlaces y la relación entre ambas magnitudes viene dada por la ecuación de Karplus (Wijmenga and van Buuren, 1998).

Una de las aplicaciones más habituales del cálculo de constantes de acoplamiento vecinales protón-protón en el campo de los ácidos nucleicos, es la determinación de la conformación del azúcar. Dado que los ángulos de torsión protón-protón están relacionados con los ángulos de torsión endocíclicos, se puede establecer una relación entre las constantes de acoplamiento de cada par de protones vecinales individuales y los parámetros de pseudorrotación (el ángulo de fase (P) y la amplitud de ángulo de pseudorrotación (ϕ_m)). De este modo se obtiene un conjunto de constantes de acoplamiento para diferentes valores de P y ϕ_m.

c) NOE (Nuclear Overhausser Effect).

El efecto Overhausser nuclear (NOE) se basa en mecanismos de relajación nuclear cruzada entre dipolos magnéticos y es, tanto más efectivo, cuanto menor es la distancia entre ellos. El efecto NOE en macromoléculas se suele medir mediante el experimento NOESY. La intensidad de los picos de correlación cruzada en el experimento NOESY es inversamente proporcional a la sexta potencia de la distancia entre los espines. Esta fuerte dependencia con la distancia provoca que sólo se observen NOEs entre protones cercanos en el espacio (aproximadamente < 6Å). La intensidad del NOE depende también de la movilidad de la molécula y del tiempo durante el cual se permite la transferencia de magnetización en el experimento NOESY (tiempo de mezcla). Además, la intensidad del NOE entre dos espines no depende solo de la posición relativa entre ambos, sino también de la posición con respecto a otros núcleos cercanos. Este efecto, llamado de difusión de espín, complica notablemente la interpretación cuantitativa del efecto NOE. En la práctica, habitualmente se lleva a cabo una interpretación cualitativa, que consiste en clasificar la intensidad del NOE en fuerte, medio y débil, y asignarles un valor de restricción de distancia para los que usualmente se consideran los intervalos 1.8-3.0, 1.8-4.0 y 1.8-5.0 Å, respectivamente. Para obtener valores de restricciones de distancia más precisos se recurre a métodos matemáticos más complejos, como el de matriz completa de relajación (Borgias and James, 1990).

2.1.2 Determinación estructural de ácidos nucleicos mediante RMN

La determinación estructural de una biomolécula en disolución consiste en obtener las coordenadas atómicas compatibles con las restricciones experimentales obtenidas por RMN. Por lo tanto, la determinación de la estructura implica la obtención del conjunto de estructuras compatibles con la información experimental. Por este motivo, las estructura de biomoléculas determinadas por RMN se muestran como una superposición de varias estructuras, que se suelen denominar *ensemble*.

Para calcular las estructuras que mejor cumplen las restricciones experimentales impuestas se aplican diversos algoritmos de cálculo. Hoy en día, el método más usado es el de dinámica molecular restringida. Este método consiste en integrar las ecuaciones de movimiento de Newton para cada átomo del sistema. La interacción entre los átomos del sistema se describe mediante un potencial fenomenológico cuya parametrización se ha ido desarrollando continuamente durante las últimas décadas. En la determinación de estructuras a partir de datos de RMN el potencial considera términos adicionales que representan la información experimental. El potencial experimental será tanto menor, cuanto mejor se cumplan las restricciones experimentales. Existen diversas implementaciones de estos métodos de cálculo. Una de las más usadas en ácidos nucleicos es la del paquete AMBER (Case et al., 2002; Cornell et al., 1995).

2.2 Pinzas ópticas

A diferencia de otras características físicas, las propiedades mecánicas de las moléculas biológicas han sido poco estudiadas hasta ahora. Las pinzas ópticas son una técnica con la cual es posible realizar un análisis mecánico utilizando una sola molécula como objeto de estudio. La posibilidad de analizar una sola molécula

en cada experimento tiene una serie de ventajas con respecto al análisis del efecto global generado por un grupo de moléculas que se realiza en técnicas como RMN, cristalografía de rayos X o microscopía electrónica. Por un lado permite obtener información sobre estados intermedios de las entidades estudiadas y, por otro, hace posible el estudio de la dinámica asociada a las fluctuaciones térmicas que se dan en el contexto celular en el que se encuentran las macromoléculas biológicas. En contraposición, la técnica de pinzas ópticas no permite obtener información acerca de la posición exacta de los átomos de la molécula en el espacio, lo cual solo es posible con las técnicas anteriormente mencionadas.

Algunas propiedades mecánicas, como la elasticidad o la tensión asociada a cambios conformacionales de los ácidos nucleicos, se miden en el rango de las unidades o las decenas de piconewtons (pN). De igual modo, las fuerzas producidas por radiación electromagnética sobre un objeto físico (fuerzas ópticas) son generalmente del orden de los pN. Por ello, el empleo de las pinzas ópticas u otras técnicas basadas en el uso de fuerzas ópticas, resulta ideal para estudiar ciertas propiedades mecánicas de los ácidos nucleicos. Las pinzas ópticas son un instrumento capaz de atrapar partículas de un tamaño en la escala de los nanómetros o los micrómetros y de medir las fuerzas que actúan sobre ellas.

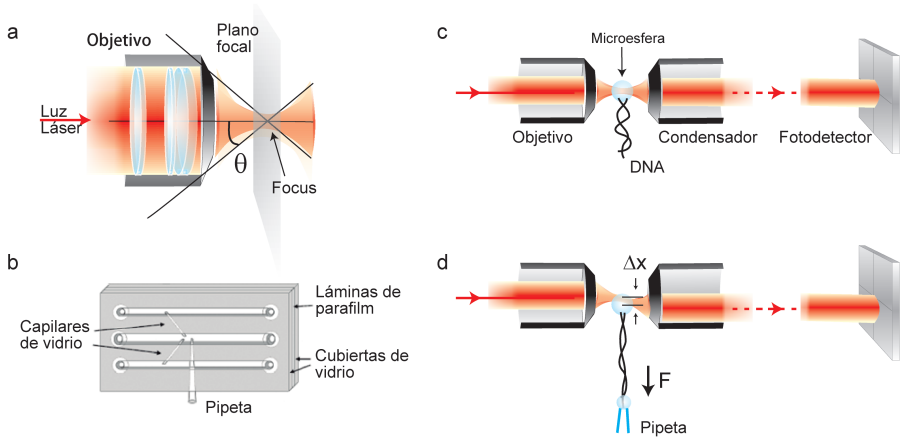


Figura 7. *(a) Representación del objetivo de un microscopio y el ángulo de apertura, θ, que forman los rayos del láser cuando se hacen converger en un foco. (b) Esquema de una cámara de fluidos en el que se designan sus diferentes componentes. (c) Esquema mostrando la configuración del objetivo frente a las lentes condensadoras que recogen la luz a la salida de la trampa óptica y un fotodetector para analizar la luz desviada. (d) Esquema mostrando la manera de medir una determinada fuerza registrando la desviación de la luz en el fotodetector. Figura reproducida con permiso de la referencia (Arias-Gonzalez, 2013).*

La luz puede ser considerada como un chorro de fotones. La interacción de la luz con una partícula provoca una variación en el momento de los fotones y como consecuencia de ello se originan fuerzas que afectan a la partícula irradiada. Las pinzas ópticas se basan en el uso de una fuente de luz láser IR focalizado por un objetivo y unas lentes condensadoras (Figura 7a) que generan una distribución de fuerzas que permite atrapar la partícula irradiada en lo que se denomina una trampa óptica. La formación de la trampa óptica tiene lugar en la cámara de fluidos cuyo diseño permite la realización de los experimentos de estiramiento y relajación así como el flujo de disolventes y muestra (Figura 7b.). Para medir las fuerzas externas generadas sobre la partícula atrapada se registra la distribución de la luz desviada de la trampa óptica mediante el uso de fotodetectores (Figura 7c y 7d). Las medidas de estas fuerzas se basan en el hecho de que la fuerza

externa que actúa sobre objetos atrapados en una trampa óptica es aproximadamente proporcional al desplazamiento generado, Δx o dicho de otra forma $F = k\Delta x$, donde k es la rigidez de la trampa óptica, un parámetro que caracteriza el elemento de unión invisible que une la partícula con la trampa óptica.

En los experimentos de pinzas ópticas para estudiar ácidos nucleicos, una molécula de DNA o RNA se ancla por sus extremos a dos bolas dieléctricas de poliestireno o de sílice, una atrapada mediante una trampa óptica y la otra mediante succión con una micropipeta (Figura 7d) o mediante otra trampa óptica. Moviendo la micropipeta en relación a la trampa óptica o una trampa óptica con respecto a la otra, el polímero puede ser estirado o relajado.

La respuesta mecánica de un polímero como es el caso de los ácidos nucleicos se puede caracterizar por su curva de fuerza-extensión (Figura 8). En estas curvas se pueden distinguir varios tipos de regímenes de elasticidad. En primer lugar el régimen de elasticidad entrópico, en el cual la molécula es estirada hasta alcanzar su longitud de contorno, lo que ocurre por debajo de los 5 pN. La aplicación de fuerzas mayores produce un régimen de elasticidad entálpico o intrínseco, en el cual la molécula se estira por encima de su superficie de contorno (valores menores que 60 pN). A partir de valores de fuerza superiores a aproximadamente 65 pN la molécula sufre una transición estructural hacia un estado casi totalmente desplegado.

Esta transición se denomina transición de sobreestiramiento u *overstretching* en la cual la molécula se estira hasta 1.7 veces su longitud de contorno en un rango muy pequeño de fuerzas (aproximadamente 2 pN). Las bases dejan de estar apiladas y las hebras pueden seguir unidas por apareamiento de bases parcial o pueden separarse dando lugar a un estado desnaturalizado. A partir del ajuste de las curvas de fuerza-extensión a modelos físicos que explican el comportamiento elástico de polímeros semiflexibles, es posible obtener valores para diferentes parámetros mecánicos.

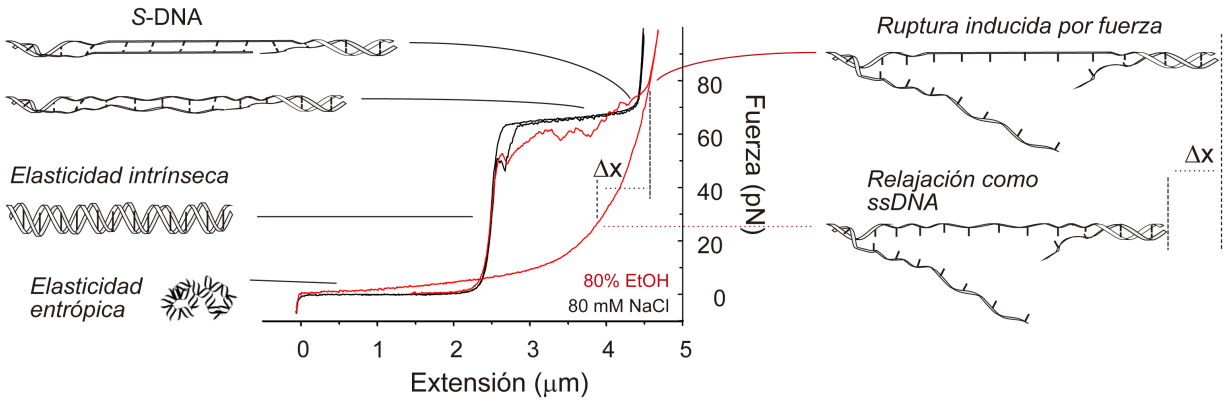


Figura 8. *Curva de fuerza extensión obtenida mediante el estiramiento (línea negra) y la posterior relajación (línea roja) de una molécula de DNA de doble cadena. Los esquemas laterales ilustran lo que le ocurre a la molécula en función de la fuerza de estiramiento o relajación aplicada. Figura reproducida con permiso de la referencia (Arias-Gonzalez, 2014).*

A la hora de interpretar los resultados obtenidos de experimentos con pinzas ópticas hay que considerar que las moléculas de DNA y RNA en disolución son estructuras dinámicas que están sujetas a fluctuaciones térmicas. Por tanto, a parte de la naturaleza puramente mecánica y electrostática de los

parámetros elásticos medidos, existe una contribución estadística que es inherente al sistema y que lleva consigo un significado biológico. Por ello, los valores de los parámetros estudiados en los experimentos de pinzas ópticas están representados por un valor medio obtenido de un número lo más elevado posible de experimentos, y por la anchura de la distribución de valores, que viene dada normalmente por la desviación estándar en distribuciones Gaussianas.

2.3 Espectrometría de masas

Una de los aspectos fundamentales para la determinación de la estructura de ácidos nucleicos es conocer el número de moléculas que intervienen en la formación de la estructura objeto de estudio. Las moléculas de DNA y RNA pueden formar estructuras intramoleculares, en las que interviene una solo molécula o intermoleculares en las que interviene más de una molécula. Así, es frecuente encontrar estructuras formadas por dos o más moléculas de DNA o RNA que se asocian entre sí. Éste es el caso de los dúplex tipo A o tipo B en los que dos hebras de RNA o DNA se asocian para formar la estructuras diméricas. Del mismo modo, se pueden formar trímeros, tetrámeros, etc.

La espectrometría de masas es una técnica de enorme eficacia en la determinación de la molecularidad o estado de oligomerización de estructuras de ácidos nucleicos. Esta técnica se basa en la separación de los componentes de una mezcla atendiendo a su diferente relación masa/carga. En todo espectrómetro de masas se pueden distinguir tres componentes: la fuente de ionización, el analizador de masas y el detector. Existen una amplia variedad de métodos de ionización y su elección depende del tipo de muestra que se desea analizar y de la eficiencia en la ionización de los componentes de la muestra. El desarrollo de métodos de ionización como el electrospray (ESI) ha sido clave para la aplicación de la espectrometría de masas al estudio de biomoléculas. Este sistema de ionización permite transferir el analito al espectrómetro de masas sin apenas producir fragmentación del analito. De este modo, es posible analizar biomoléculas cuya estructura está estabilizada por interacciones no covalentes, como es el caso de complejos de proteínas (Loo, 2000) o estructuras de ácidos nucleicos (Beck et al., 2001; Hofstadler and Griffey, 2001).

La resolución del espectrómetro tan solo influye en la complejidad de las muestras que pueden ser analizadas con un único espectro. Por otro lado, el nivel de resolución determina si es posible obtener la distribución isotópica de las diferentes especies de la muestra. Esto puede ser de gran ayuda para asignar la carga de un pico ya que los isótopos están separados por 1 Da, de manera que en la escala m/z el espaciado entre picos consecutivos de la distribución isotópica es igual al inverso de la carga: 1/z. Una vez que la carga es conocida, la masa se obtiene multiplicando el valor m/z del pico por la carga.

Para preservar la estructura de los ácidos nucleicos, es preciso utilizar disoluciones que contengan cationes monovalentes. Esto supone un problema a la hora de utilizar la espectrometría de masas con ionización por electrospray ya que el mecanismo de generación de los iones (Amad et al., 2000; Cole, 2000; Kebarle, 2000) provoca la condensación de los cationes monovalentes en las moléculas de ácido nucleico. La presencia de mínimas cantidades de cationes de sodio o potasio resulta en la detección de una amplia distribución de estequiometrias de aductos en el DNA. Por ello, el aspecto clave en la preparación de las muestras es formar la estructura deseada en un contexto de ausencia total de cationes sodio y potasio. Esto

se logra normalmente usando acetato amónico en lugar de NaCl o KCl y realizando los ajustes de pH necesarios con ácido acético y amoniaco. La imposibilidad de realizar estudios en presencia de las sales presentes en condiciones fisiológicas hace que sea necesario confirmar mediante otras técnicas si la estructura que adopta una determinada secuencia es igual en presencia de cationes amonio y de cationes sodio o potasio.

2.4 Dicroísmo circular

El dicroísmo circular (DC) es una técnica que permite obtener información sobre la estructura secundaria de macromoléculas biológicas. Un haz de luz polarizado en un plano puede considerarse como constituido por dos componentes polarizados circularmente, uno a la derecha y el otro a la izquierda. La interacción de la luz polarizada en un plano con los centros quirales de la molécula irradiada provoca una absorción diferencial de los dos componentes de la luz polarizada circularmente con la que se irradia. La diferente absorción de los dos componentes de la luz circularmente polarizada provoca que el vector suma de los vectores eléctricos de cada componente (E_L + E_R) describa una elipse. El DC se define mediante la relación entre el eje semimayor y semimenor de dicha elipse. Esta relación es la tangente del ángulo θ, conocido como elipticidad. En los experimentos de DC se registra la diferencia entre la absorción de la luz polarizada circularmente hacia a la izquierda y hacia la derecha ($\Delta A = A_L - A_R$), que se relaciona con la elipticidad (θ) según la siguiente expresión: $\theta = 32{,}98.\Delta A$. La elipticidad varía con la longitud de onda, por lo que se pueden obtener espectros en los que se representa la elipticidad versus la longitud de onda. Otro tipo de experimentos consisten en registrar la variación de elipticidad con la temperatura manteniendo una longitud de onda fija. En este tipo de experimentos se obtienen curvas de desnaturalización que permiten determinar la temperatura de desnaturalización de la estructura estudiada.

Las moléculas analizadas por DC deben absorber radiación UV–visible y ser quirales. El carácter quiral de una molécula puede ser debido a la presencia de elementos intrínsecamente quirales o a la localización de cromóforos en un entorno asimétrico. El DC de los ácidos nucleicos tiene su origen en la asimetría de los azúcares del esqueleto azúcar-fosfato y en el plegamiento que adopta la cadena de nucleótidos. Los espectros de DC de ácidos nucleicos se suelen realizar variando la longitud de onda entre 200 y 320 nm. En este intervalo de longitudes de onda se producen las transiciones electrónicas de las bases nitrogenadas. La posición de las bandas positivas y negativas del espectro es indicativo del tipo de estructura secundaria que está formando la molécula de DNA o RNA (Vorlickova et al., 2012).

3 Ácidos nucleicos en telómeros y centrómeros

Los telómeros y los centrómeros son las regiones del cromosoma implicadas en la protección y en la segregación cromosómica, respectivamente. Se trata de regiones heterocromáticas, en las que el DNA constituyente presenta una mayor cantidad de elementos repetitivos, tales como repeticiones en tándem (DNA satélite) y elementos transponibles. Esta naturaleza repetitiva otorga a los telómeros y centrómeros características particulares desde el punto de vista estructural.

3.1 Biología y estructura de los telómeros

Los telómeros son los complejos de ácidos nucleicos y proteínas que se encuentran en los extremos de los cromosomas lineales de los eucariotas (Figura 9a). La función de los telómeros es impedir que los finales de los cromosomas sean identificados por la célula como roturas de DNA de doble cadena y sean inoportunamente procesados por la maquinaria de reparación de DNA de la célula, lo que provocaría la fusión de unos cromosomas con otros a través de sus extremos.

El DNA de los telómeros está formado por secuencias de nucleótidos que se repiten. La unidad de repetición puede variar un poco dependiendo de la especie. En los telómeros de todos los vertebrados, la secuencia repetida en tándem es TTAGGG y su extensión es de 5-8 kilobases (kb). El extremo 3' de cada telómero esta formado por la cadena sencilla rica en guaninas que sobre sale unos 100-200 nucleótidos (Figure 9b).

Algunas proteínas del telómero se asocian directamente al DNA, ya sea al DNA de doble cadena como es el caso de TRF1 y TRF2 o al extremo 3'-monohebra como ocurre con POT1. En la heterocromatina telomérica existen otras proteínas que se encuentran asociadas a las anteriormente mencionadas (Figura 9b) y que completan el complejo proteico telosoma o shelterina.

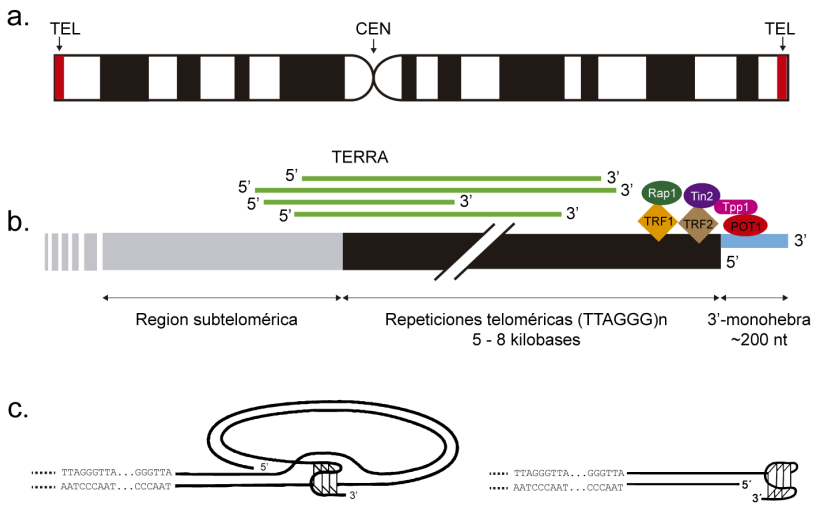


Figura 9. *(a) Esquema de un cromosoma marcando la posición de los telómeros (rojo) y del centrómero. (b) Esquema de la composición de un telómero en el que se muestran las proteínas teloméricas, diferentes moléculas de TERRA y las distintas regiones de DNA telomérico y subtelomérico con su correspondiente longitud aproximada. (c) Extremo 3'-monohebra formando un cuádruplex de guanina (derecha) y un cuádruplex de guanina al final de un T-loop (izquierda).*

Se ha descubierto recientemente que el DNA telomérico es transcrito por la enzima RNA polimerasa II dando lugar a moléculas de RNA telomérico (Azzalin et al., 2007; Schoeftner and Blasco, 2007). La hebra rica en citosinas del DNA telomérico sirve como molde para la transcripción del RNA telomérico denominado TERRA (*Telomeric repeat-containing RNA*). La transcripción tiene su origen en la región subtelomérica (Nergadze et al., 2009; Pfeiffer and Lingner, 2012) por lo que las moléculas de TERRA contienen secuencias subteloméricas y terminan con un número variable de repeticiones ricas en guaninas (5'-UUAGGG-3' en el caso de los vertebrados). Además de TERRA, existen otros RNAs que completan el transcriptoma del telómero. Así, en las levaduras se han identificado también el RNA de la hebra

complementaria a TERRA, que recibe el nombre de ARIA, así como RNAs subteloméricos de ambas hebras, conocidos como ARRET y αARRET (Bah et al., 2012; Greenwood and Cooper, 2012).

TERRA es un RNA no codificante que forma parte de la heterocromatina telomérica (Figura 9b). Desde su descubrimiento se han sugerido diversas funciones para TERRA. Hasta el momento se sabe que los niveles de TERRA tienen influencia sobre la regulación de la longitud de los telómeros (Cusanelli et al., 2013; Pfeiffer and Lingner, 2012; Redon et al., 2010; Redon et al., 2013). Dicha regulación tiene lugar a través de diversos mecanismos que implican la interacción de TERRA con proteínas del telómero y con enzimas que alteran la longitud del mismo, tales como la telomerasa o la exonucleasa 1. Por otro lado, se ha observado que TERRA interviene en los cambios de composición proteica que ocurren en los telómeros durante el ciclo celular (Flynn et al., 2011) o durante los procesos de senescencia (Porro et al., 2014). Además, TERRA parece estar implicado también en la movilidad de los telómeros (Arora et al., 2012). Por último, el aumento en los niveles de TERRA en la célula se ha asociado con procesos moleculares propios de enfermedades como el cáncer (Episkopou et al., 2014; Ng et al., 2009) o el síndrome de inestabilidad del centrómero y anomalía facial (ICF) (Yehezkel et al., 2008).

La abundancia de guaninas en el DNA y el RNA telomérico hace posible que se formen estructuras no canónicas del tipo cuádruplex de guanina (sección 1.2.3). En el extremo 3'-monohebra del DNA telomérico la formación de éstas estructuras no canónicas de DNA está más facilitada debido a la ausencia de la hebra complementaria con la que formar una estructura B-DNA (Figura 9c). Lo mismo ocurre con las moléculas de TERRA, puesto que también se trata de moléculas de cadena sencilla con un alto contenido en guaninas. La formación de cuádruplexes de guanina de DNA y de RNA telomérico ha sido confirmada mediante numerosos estudios *in vitro* en los que se analizan secuencias cortas que permiten la caracterización estructural mediante técnicas biofísicas. Así, se ha observado que el DNA telomérico es capaz de adoptar diferentes conformaciones de cuádruplex de guanina en función de las condiciones experimentales, mientras que el RNA telomérico se pliega siempre en la conformación de quadruplex paralelo. Sin embargo existe escasa información sobre la estructura de secuencias de longitud más próxima a la que tienen estas moléculas en la célula.

La formación de cuádruplex de guanina de DNA *in vivo* ha sido demostrada recientemente en células humanas (Biffi et al., 2013) observándose, como era de esperar, un predominio de estas estructuras en las regiones teloméricas. La existencia de los cuádruplexes de guanina de RNA en la célula está todavía por determinar mediante métodos directos. Sin embargo, estudios recientes han mostrado que oligonucleótidos de RNA telomérico adoptan una estructura tipo cuádruplex de guanina paralelo dentro de la célula (Xu et al., 2010).

El hecho de que el RNA telomérico se localice en el telómero (Azzalin et al., 2007; Luke et al., 2008; Schoeftner and Blasco, 2007) sugiere la posible formación de estructuras híbridas. Recientemente se ha demostrado la formación *in vitro* de cuádruplex híbridos de DNA y RNA teloméricos que, además, presentan una resistencia superior a la hidrólisis por nucleasas que los correspondientes cuádruplexes de guanina no

híbridos (Xu et al., 2014; Xu et al., 2009). Estos resultados sugieren que estas estructuras híbridas podrían estar implicadas en la protección de los finales del cromosoma.

3.1.1 Telómero y cáncer

Durante las últimas décadas se han encontrado numerosas evidencias que indican que los telómeros tienen un papel fundamental en el control de la supervivencia de la célula. El acortamiento de los telómeros que tiene lugar durante la replicación del DNA desemboca en la muerte de la célula tras un número determinado de divisiones. Esto no ocurre en aquellas células que consiguen mantener la longitud de los telómeros, como en el caso de la mayoría de las células tumorales.

Para contrarrestar el progresivo acortamiento de los telómeros, las células pueden emplear dos mecanismos. El primero de ellos se basa en la acción de la enzima telomerasa, que mediante su actividad transcriptasa inversa, es capaz de añadir nucleótidos al extremo 3' del cromosoma utilizando un RNA como molde. Este mecanismo está activo en el 85-90 % de las células tumorales. En el 10-15 % restante el mecanismo empleado para mantener la longitud de los telómeros se denomina ALT, y se basa en la extensión de los telómeros mediante eventos de recombinación.

Debido a la sobreexpresión de la telomerasa en las células tumorales, esta enzima se ha convertido en una importante diana terapéutica. Se han diseñado diversas estrategias para inhibir la acción de la telomerasa y así evitar la inmortalidad de las células malignas (Olaussen et al., 2006). Una de estas estrategias es la de utilizar compuestos que estabilicen estructuras tipo cuádruplex de guanina en el DNA o RNA telomérico. La estabilización de estas estructuras mediante su interacción con ligandos conduce a la disrupción de la heterocromatina telomérica y en consecuencia provoca la muerte celular. Son diversos los mecanismos que se han propuesto para explicar la acción de los ligandos que estabilizan el cuádruplex de guanina en secuencias teloméricas. Por un lado se ha propuesto que la estabilización de los cuádruplex de guanina impide que el extremo 3'-monohebra pueda interaccionar con el RNA molde de la telomerasa y en consecuencia se evita la elongación de los telómeros (Sun et al., 1997; Zahler et al., 1991). Por otro lado se ha observado que la presencia de estos ligandos activa la respuesta a daño de DNA (Rodriguez et al., 2008) y produce fusiones entre finales de cromosomas en metafase (Incles et al., 2004) lo cual puede estar relacionado con el desplazamiento de la proteína hPOT1 o de la telomerasa del extremo 3' del cromosoma. Posiblemente, son varios los mecanismos que entran en juego como consecuencia de la estabilización de los cuádruplex de guanina en el telómero. Prueba de ello es que existen compuestos, que muestran que la estabilización del cuádruplex de guanina en el telómero es una estrategia válida, no solo en células en las que la telomerasa se sobreexpresa sino también para células que utilizan el mecanismo ALT de elongación de los telómeros (Pennarun et al., 2005; Riou et al., 2002).

Los compuestos que se han diseñado hasta el momento para unir selectivamente y estabilizar cuádruplex de guanina comparten tres características principales: (i) Tienen una plataforma aromática para maximizar las interacciones π con la superficie de las tétradas de guanina. (ii) Tienen carga positiva para aumentar la afinidad por los grupos fosfato del DNA. (iii) Presentan cadenas laterales que pretenden maximizar las interacciones con los surcos, los *loops*, o las bases del DNA. Son muchos los compuestos que

se han diseñado y estudiado como agentes estabilizadores de cuádruplex de guanina. Algunos ejemplos destacados de este tipo de compuestos son las acridinas BRACO-19 (Burger et al., 2005) y RHSP4 (Leonetti et al., 2008; Phatak et al., 2007; Salvati et al., 2007) o la porfirina TMPyP4 (Grand et al., 2002) cuyos potenciales de inhibición de telomerasa han sido probados en ensayos *in vivo*. Sin embargo, la inmensa mayoría de los compuestos que se han estudiado hasta el momento tienen características poco adecuadas para ser utilizadas como fármacos y tan solo un compuesto, la quarfloxina o CX-3543 (Drygin et al., 2009), ha logrado alcanzar fases clínicas.

3.2 Biología y estructura del centrómero

El centrómero es la región del cromosoma donde tiene lugar el ensamblaje del cinetocoro, una estructura multiproteíca implicada en el movimiento de los cromosomas durante la segregación cromosómica. Se puede percibir visualmente como una constricción que aparece en los cromosomas metafásicos, que recibe el nombre de constricción primaria (Figura 10).

Las proteínas asociadas a la cromatina centromérica reciben el nombre de CENPs e incluyen, entre otras, CENP-A, CENP-B, CENP-C y CENP-T. La proteína CENP-A es una variante de la histona H3 y contribuye en la formación de una cromatina especializada en los centrómeros. La cromatina centromérica está formada por bloques que contienen nucleosomas con la histona H3 intercalados con nucleosomas con CENP-A. La presencia de CENP-A en los nucleosomas se considera la marca epigenética de la cromatina centromérica.

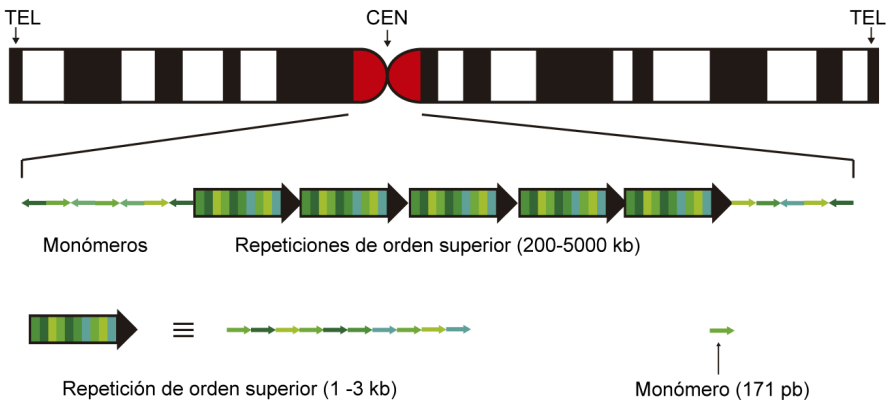


Figura 10. *Esquema de un cromosoma marcando la posición del centrómero (rojo) y de los telómeros. Mas abajo se muestra una visión extendida de la región centromérica, mostrando la repetición de unidades de orden superior flanqueadas por monómeros de la secuencia alfoide. Las unidades de orden superior están formadas por la repetición en tándem de monómeros de la secuencia alfoide, de una extensión de 171 pares de bases.*

Una característica sorprendente de los centrómeros es que las secuencias de DNA presentes en los mismos varían de unos organismos a otros e incluso entre cromosomas de un mismo organismo. Por el contrario, las proteínas centroméricas están conservadas entre especies, lo cual sugiere una base epigenética para el ensamblaje del centrómero. Sin embargo, existen argumentos que apoyan la importancia de la secuencia de DNA en la formación del centrómero como por ejemplo en la formación de centrómeros *de novo* en los cromosomas artificiales humanos (Harrington et al., 1997; Ikeno et al., 1998). Por el contrario,

otras evidencias, como la formación de centrómeros en secuencias variables dentro de una misma especie, tal y como ocurre con la formación de neocentrómeros (Choo, 2001; Warburton, 2004) respaldan mecanismos epigenéticos en la formación de la cromatina centromérica.

A pesar de la falta de identidad de secuencia, muchos centrómeros están localizados en regiones de DNA altamente repetido o DNA satélite. En humanos y en la mayoría de los primates, el DNA satélite alfoide aparece en el centrómero. La secuencia alfoide puede llegar a comprender hasta el 5% de todo el genoma y aparece en la constricción primaria de todos los cromosomas humanos. Este satélite está formado por la repetición en tándem de una secuencia de 171 pares de bases. Las sucesivas repeticiones (monómeros), comparten entre el 50 y el 70% de similitud de secuencia. La repetición de un número entero de monómeros da lugar a una unidad de repetición de orden superior que a su vez se repite, llegando a ocupar entre 250 y 5000 kilobases (Figura 10). Las repeticiones de orden superior también se diferencian entre sí pero de una forma mucho menos acusada que los monómeros que los constituyen. Flanqueando la sucesión de repeticiones de orden superior se encuentran monómeros distribuidos de forma aleatoria que conectan la zona de repeticiones homogéneas con los brazos del cromosoma (Figura 10).

El número de veces que se repite la unidad de orden superior es variable, por lo que la longitud total del satélite varía entre cromosomas homólogos y entre individuos. Por otro lado la variación en la composición del satélite puede ocurrir por la diferencia de tamaño de las unidades de repetición de orden superior. Así, por ejemplo, la unidad de repetición de orden superior puede estar formada por un número variable de monómeros. Otro nivel de variación se produce en los monómeros que constituyen la repetición de orden superior. Algunos monómeros contienen una secuencia de 17 nucleótidos rica en GCs que recibe el nombre de CENP-B box por ser el lugar de unión de la proteína CENP-B. En su defecto, otros monómeros no presentan dicha secuencia o presentan una variación de la misma, también de 17 nucleótidos y rica en GCs en la zona equivalente del monómero.

Debido al desconocimiento que se tiene hasta el momento sobre la secuencia completa de los centrómeros, son pocos los estudios estructurales relativos a secuencias centroméricas. La secuencia CENP-B box ha sido estudiada mediante RMN, revelando que la hebra rica en citosinas es capaz de formar una estructura dimérica tipo *i-motif* (Gallego et al., 1997; Gallego et al., 1999) (Ver Sección 1.2.3). Atendiendo a estos resultados se ha sugerido que la unión de la proteína CENP-B a la secuencia CENP-B box podría facilitar la formación de un *i-motif* por parte de la hebra rica en Cs (Gallego et al., 1997; Gallego et al., 1999). También se ha sugerido que CENP-B podría organizar una estructura de orden superior en el centrómero en la cual dos secuencias CENP-B box distantes se sitúan yuxtapuestas a través de interacciones proteína-proteína y DNA-proteína (Yoda et al., 1992).

Objetivos

OBJETIVOS

Esta tesis doctoral fue concebida con los siguientes objetivos:

- Explorar la capacidad que tienen las secuencias teloméricas y centroméricas de formar estructuras no canónicas de ácidos nucleicos.

- Caracterizar las estructuras no canónicas formadas por secuencias teloméricas y centroméricas mediante el empleo de distintas aproximaciones experimentales.

- Aplicar metodologías que permitan buscar compuestos que estabilicen estructuras no canónicas de ácidos nucleicos en secuencias de RNA telomérico de elevado peso molecular.

- Plantear la posible relevancia biológica de las estructuras no canónicas de los ácidos nucleícos en la función y evolución de telómeros y centrómeros.

Materiales y Métodos y Resultados

Artículo 1.

On the origin of the eukaryotic chromosome: The role of noncanonical structures in telomere evolution

Miguel Garavís, Carlos González and Alfredo Villasante

Sobre el origen del cromosoma eucariota: El papel de las estructuras no canónicas en la evolución del telómero

Miguel Garavís, Carlos González and Alfredo Villasante

Se ha conjeturado que la endosimbiosis de una α-proteobacteria (simbionte) en una arqueobacteria (huésped), provocó la invasión de los "intrones tipo II" del simbionte dentro del genoma circular del huésped. De esta forma los "intrones tipo II" dieron lugar a los intrones eucarióticos, algunos de los cuales, tras perder su capacidad de *splicing*, dieron lugar a su vez a los retrotransposones non-LTR (*non-Long Terminal Repeat*). La inestabilidad cromosómica generada en ese contexto evolutivo pudo conducir a una ruptura continua del cromosoma circular, lo cual probablemente activó los mecanismos de reparación de DNA de la célula, uno de los cuales consiste en la migración de retrotransposones no-LTR a los finales del cromosoma.

En este trabajo proponemos que la acumulación de retrotransposones con una distribución sesgada de nucleótidos G/C en ambas hebras podría haber sido seleccionada debido a que de las secuencias ricas en guanina tienen una capacidad intrínseca de formar estructuras no canónicas que protegerían el final del cromosoma (*capping*).

Para asegurar la viabilidad de la célula, los finales de los nuevos cromosomas lineales (prototelómeros) deberían poder desempeñar, tanto la función de *capping*, como la de segregación del cromosoma durante la división celular. Mas tarde, con la evolución de los centrómeros en la región subtelomérica, se diferenciarían los telómeros y los centrómeros como estructuras especializadas en la realización de las funciones de *capping* y segregación, respectivamente.

En la mayoría de los cromosomas eucariotas, las secuencias de DNA telomérico son repeticiones de una secuencia corta de unos pocos nucleótidos que se repiten en tándem y que terminan en un extremo 3' monohebra. En muchos organismos la enzima telomerasa mantiene la longitud de los telómeros mediante la síntesis de repeticiones teloméricas en el final del cromosoma. Existen evidencias que sugieren que el motivo TTAGGG es la secuencia telomérica ancestral de los eucariotas. Además, estudios estructurales han mostrado que esta secuencia es capaz de plegarse formando estructuras tipo cuádruplex de guanina, siendo además la que lo hace con mayor facilidad *in vitro*. Algunos organismos presentan otros motivos de repetición en el telómero que no forman con tanta facilidad estructuras tipo cuádruplex de guanina. En estos casos, además, la actividad de la telomerasa es mas reducida, y por tanto son necesarias otras estrategias para mantener la longitud del telómero. Así, las secuencias teloméricas de algunos artrópodos están compuestas por repeticiones de una secuencia corta y por retroelementos que se intercalan entre las repeticiones de tal manera que se conserve el desequilibrio de G/C en ambas hebras. Algunas levaduras presentan secuencias teloméricas variables como resultado de la baja procesividad de sus telomerasas. Estas secuencias tienen una baja tendencia a formar cuádruplex de guanina. Sin embargo, se ha observado que en estos organismos, las proteínas de unión al telómero evolucionan rápidamente, lo que sugiere que en estos

casos se puede estar usando un sistema ancestral de protección del cromosoma en el que las proteínas de unión a DNA monohebra facilitan el plegamiento del DNA en estructuras tipo cuádruplex.

Algunos organismos carecen de telomerasa, de manera que usan otros mecanismos para mantener sus telómeros. Tal es el caso de los dípteros, que mantienen la longitud de sus telómeros por retrotransposición de elementos móviles y procesos de recombinación, un mecanismo que recuerda al que pudo tener lugar durante la formación de los prototelómeros. Algunas levaduras también carecen de telomerasa y elongan sus telómeros mediante amplificaciones y recombinaciones de secuencias subteloméricas y rDNA. En ambos tipos de organismos, se mantiene un desequilibrio de G/C entre las dos hebras de la región telomérica, lo que apoya de nuevo la hipótesis de que el *capping* en el telómero depende de la formación de estructuras no canónicas basadas en interacciones G:G.

Aportación personal al trabajo: Planteamiento, discusión y revisión del manuscrito.

Genome Biology and Evolution

On the Origin of the Eukaryotic Chromosome: The Role of Noncanonical DNA Structures in Telomere Evolution

Miguel Garavís[1,2], Carlos González[2], and Alfredo Villasante[1,*]

[1]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Nicolás Cabrera 1, 28049 Madrid, Spain

[2]Instituto de Química Física Rocasolano, CSIC, C/Serrano 119, 28006 Madrid, Spain

*Corresponding author: E-mail: avillasante@cbm.uam.es.

Associate editor: Bill Martin

Accepted: May 16, 2013

Abstract

The transition of an ancestral circular genome to multiple linear chromosomes was crucial for eukaryogenesis because it allowed rapid adaptive evolution through aneuploidy. Here, we propose that the ends of nascent linear chromosomes should have had a dual function in chromosome end protection (capping) and chromosome segregation to give rise to the "proto-telomeres." Later on, proper centromeres evolved at subtelomeric regions. We also propose that both noncanonical structures based on guanine–guanine interactions and the end-protection proteins recruited by the emergent telomeric heterochromatin have been required for telomere maintenance through evolution. We further suggest that the origin of *Drosophila* telomeres may be reminiscent of how the first telomeres arose.

Key words: telomeres, centromeres, G-quadruplexes, non-LTR retrotransposons.

SMBE

The Necessity of Both End Protection and Segregation Functions at the End of Nascent Linear Chromosomes

It has been hypothesized that after the endosymbiosis of an α-proteobacteria into an archaebacterial host, the massive invasion of the symbiont's mobile group II introns into the circular genome of the host gave rise to spliceosomal introns (Koonin 2006) and that this was the driving force for the origin of the nucleus (fig. 1). The invention of the nuclear membrane was necessary to physically separate slow splicing from fast translation (Martin and Koonin 2006) (fig. 1). It has also been proposed that this invasion eventually lead to the origin of linear eukaryotic chromosomes (Villasante, Abad, et al. 2007; Villasante, Méndez-Lago, et al. 2007). The host's tolerance to the mobile element invasion and to other eukaryotic innovations could be facilitated by a low effective population size and the consequent weak purifying selection (Lynch and Conery 2003; Lynch 2007; Koonin 2011).

Mobile group II introns are retroelements of eubacterial origin that contain a catalytic RNA and a multifunctional protein with reverse transcriptase (RT) activity (Lambowitz and Zimmerly 2004). These retroelements are thought to be the ancestors of both spliceosomal introns and nonlong terminal repeat (non-LTR) retrotransposons (Sharp 1985). The evolutionary relationship between group II introns and non-LTR-retrotransposons is based on the similarity of their RT sequences (Xiong and Eickbush 1990; Blocker et al. 2005) and retrotransposition mechanisms (Luan et al. 1993; Zimmerly et al. 1995). After the initial proliferation of group II introns within the protoeukaryotic nuclear genome, their RNA domains degenerated and evolved into spliceosomal snRNAs that functioned in *trans* in a common splicing apparatus (Sharp 1991; Mohr et al. 2010). Although most group II introns evolved as eukaryotic introns, some lost their splicing capability and gave rise to non-LTR-retrotransposons.

It is likely that the continuous breakage of the presumed circular chromosome activated all the mechanisms of DNA repair, including the one mediated by non-LTR retrotransposons (Moore and Haber 1996; Morrish et al. 2002). In this evolutionary scenario, it has been hypothesized that the repetitive capture of non-LTR retrotransposons, with a G/C strand bias, at the ends of DNA double-strand breaks (DSBs) could have eventually resulted in end protection (capping), instead of repair, giving rise to the "proto-telomeres" of the first linear chromosomes (fig. 1) (Villasante, Abad, et al. 2007). The biased distribution of guanine and cytosine between the two strands could have been selected because G-rich

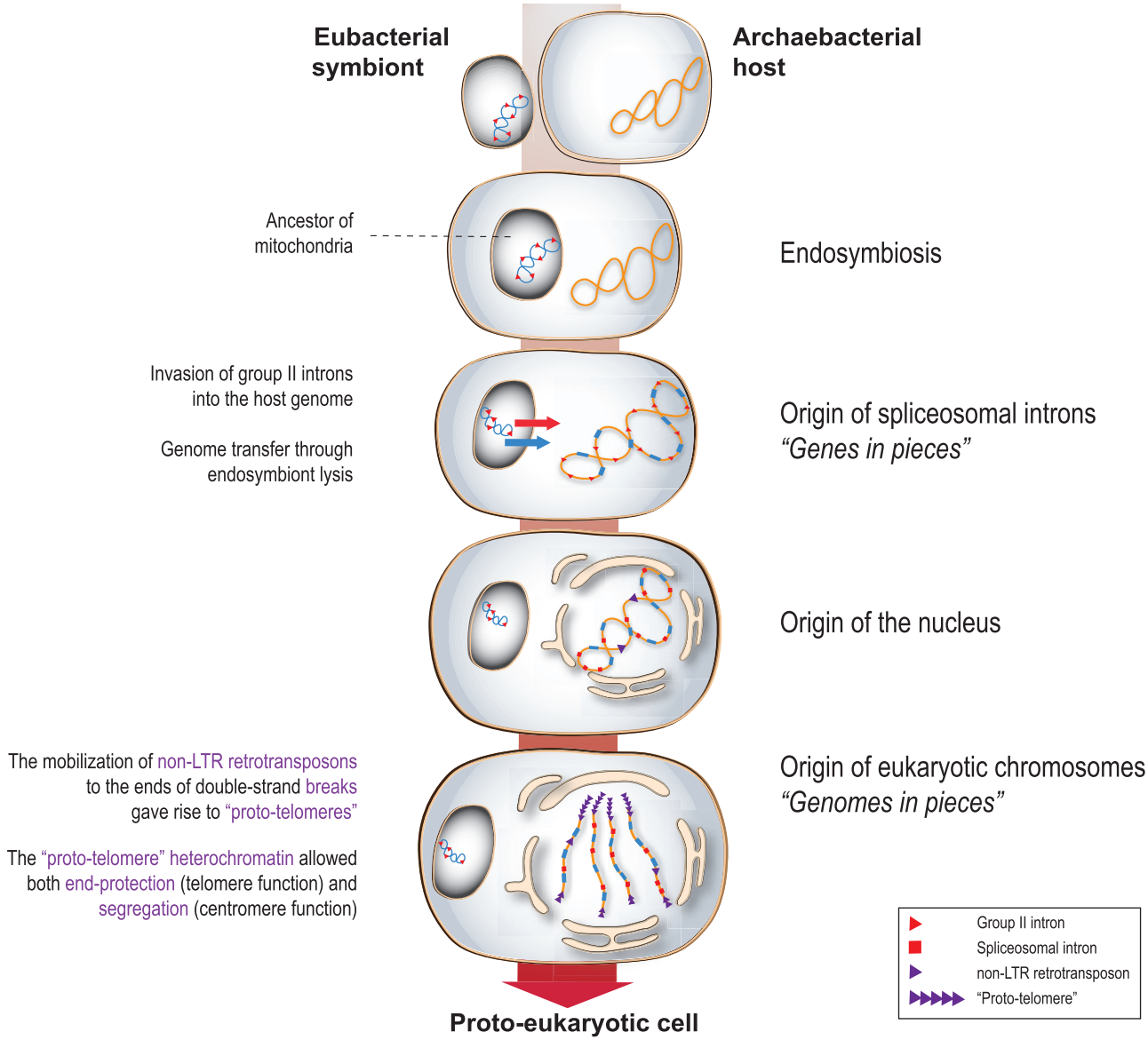


Fig. 1.—Schematic representation of a possible evolutionary scenario for the origin of eukaryotic chromosomes. The scheme shows the origin of spliceosomal introns and the origin of the nucleus as hypothesized (Koonin 2006; Martin and Koonin 2006) and also the proposed origin of the ancestral "proto-telomere" with capping and segregation properties. The eubacterial genome appears in blue and the archaebacterial genome in orange.

sequences have the intrinsic capacity to fold into noncanonical secondary structures that were utilized for capping or sequestering chromosome ends (Villasante, Abad, et al. 2007; Villasante, Méndez-Lago, et al. 2007). In addition, the iterative transposition generated the first terminal repeats and that also allowed the elongation of chromosome ends by the existing mechanisms of homologous recombination (de Lange 2004). As will be described later, a similar situation occurs in *Drosophila*, where telomeres are maintained by retrotransposition of G/C strand-biased non-LTR elements and by terminal recombination (Mason et al. 2008; Villasante et al. 2008).

The universal 5′–3′ nuclease-mediated cleavage of DNA ends created 3′-single-stranded overhangs that were coated by an abundant single-stranded DNA (ssDNA) -binding protein (replication protein A or RPA) via its characteristic oligosaccharide/oligonucleotide-binding fold (OB fold) domains. This nonsequence-specific ssDNA-binding protein is required for multiple processes such DNA replication, DNA repair, and DNA damage signaling. Therefore, to protect the DNA ends and to avoid their repair, a specialized sequence-specific ssDNA-binding protein should evolve to coat the 3′-G-rich overhangs (Gelinas et al. 2009) and to promote their folding into non-B DNA conformations, which could be (but not

necessarily) quadruplex-like structures. Recent studies have shown that molecular crowding (Heddi and Phan 2011; Xu et al. 2011) or high viscosity conditions (Lannan et al. 2012) stabilize G-quadruplexes, suggesting that cell environment may facilitate the formation of quadruplex-like structures.

On the other hand, a cell-cycle-regulated switch of those ssDNA-binding proteins was needed to re-establish telomere capping after DNA replication, and now it is known that telomeric repeat-containing RNA (TERRA) contributes to induce that switch (Flynn et al. 2011). TERRA is also required to organize a special chromatin structure: the telomeric heterochromatin (Azzalin et al. 2007; Schoeftner and Blasco 2008; Deng et al. 2009; Shpiz et al. 2011).

Here, it is fundamental to notice that "proto-telomeres" with dual function in capping and segregation were required to ensure accurate inheritance of the first linear eukaryotic chromosomes (fig. 1). Thus, the formation of the first heterochromatin at nascent chromosome ends should have facilitated both the recruitment of end-protection proteins and the attachment of spindle microtubules, most likely by means of ribonucleoprotein complexes (Villasante, Abad, et al. 2007; Villasante, Méndez-Lago, et al. 2007). Later on in eukaryogenesis, a mature segregation function evolved at subtelomeric regions. The mechanism of unequal exchange and gene conversion led inevitably to the divergence of the internal subtelomeric repeats, and the strand asymmetry of the repeats provided the potential to form the sequence-independent secondary structures that gave rise to the centromeres (Villasante, Abad, et al. 2007; Villasante, Méndez-Lago, et al. 2007). In this scenario, the recurrent appearance of unstable dicentric chromosomes, through the formation of new centromeres from telomeres, provided an additional mechanism of genome fragmentation. The birth of multiple eukaryotic linear chromosomes was the key innovation that allowed adaptive evolution by means of transient aneuploidy (chromosomal duplications) (Chen et al. 2012; Yona et al. 2012).

If primitive centromeres evolved from "proto-telomeres," it would be reasonable to expect that telomeric regions may also have some centromere-like properties. Indeed, there are already results that seem to support this assertion.

1) In *Schizosaccharomyces pombe*, deletion of an endogenous centromere leads to neocentromere formation at subtelomeric regions (Ishii et al. 2008), and it has been shown that their centromeric and subtelomeric DNA sequences must possess particular features that promote the incorporation of the centromere-specific histone 3 variant (CENP-A) (Choi et al. 2012).

2) In *Drosophila melanogaster*, the centromere of the Y chromosome contains a large array of *HeT-A*- and *TART*-derived telomeric retrotransposons (Agudo et al. 1999), and the sequence of this satellite DNA has revealed that this centromeric region evolved from a telomere (Méndez-Lago et al. 2009). Furthermore, overexpression of the *Drosophila* CENP-A induces preferential formation of neocentromeres near telomeres (Heun et al. 2006; Olszak et al. 2011).

3) In some plants and animals, neocentromere activity appears at subtelomeric heterochromatin during meiosis (reviewed in Puertas and Villasante 2013).

4) The evolutionary history of chromosome 3 in primates shows at least three examples of telomere–centromere functional interchange (Ventura et al. 2004). Similarly, other telomere-to-centromere conversions have been described after the comparative analysis of eight mammalian genomes (Murphy et al. 2005). Because the subtelomeric repeats could have a role in these conversions, this chromosomal behavior could be due to the ancestral centromeric competence of a telomeric region.

Similarly, if primitive centromeres began at DSBs, one could wonder whether the dynamic chromatin formed around breakage sites could have centromere-like features. Here, too, there are results in favor of this consideration.

1) It has been shown that the centromeric proteins CENP-A, CENP-N, CENP-T, and CENP-U are rapidly recruited to DSBs (Zeitlin et al. 2009) and has been hypothesized that, under certain circumstances, this recruitment could generate a neocentromere (Zeitlin et al. 2009).

2) Strikingly, it had been previously noticed that several human neocentromeres were located near breakpoints and had been hypothesized that these breaks could induce the emergence of neocentromeres (Ventura et al. 2003; Marshall et al. 2008).

The previous hypothesis for the origin of the eukaryotic chromosome proposed that centromeres arose before telomeres and that probably evolved from the origin of replication region of the bacterial chromosome (Cavalier-Smith 1981). Recently, Cavalier-Smith (2010) has still suggested that centromeres arose first and has proposed that they originated from the partitioning locus, a region proximal to the bacterial origin of replication implicated in bacterial chromosome partitioning/segregation. But he did not say how the fragmented prokaryotic genome could give rise to a centromere on each nascent linear chromosome and what was the hypothetical process that led to the formation of regional centromeres containing repetitive DNA. In support of an ancestral regional centromere, a recent study in *Saccharomyces cerevisiae* has found centromere-like regions (without a specific DNA sequence) in close proximity to the native point centromere (Lefrançois et al. 2013). Because these small regions promote proper segregation, possibly through sequence-independent centromeric structures, they seem to be evolutionary remnants derived from a regional centromere rather than from a point centromere (Lefrançois et al. 2013).

To recapitulate, in this section, we have proposed that the origin of linear chromosomes (genomes in pieces) was a eukaryotic innovation generated by the mobilization of group II intron-derived retroelements as a response to endosymbiosis stress (McClintock 1984; Koonin 2011). Specifically, we have

hypothesized that the repetitive capture of G/C strand biased non-LTR retrotransposons at the ends of DSBs gave rise to "proto-telomeres," a primitive terminal heterochromatic structure with a dual function: end protection (telomeric function) and segregation (centromeric function) (fig. 1).

Noncanonical DNA Structures Based on Guanine–Guanine Interactions Seem to Have Played a Role in Telomere Origin and Evolution

In most eukaryotic chromosomes, telomere DNA sequences are arrays of short guanine-rich repetitive sequences that terminate in a 3′-single-strand G-rich overhang (150–200 nucleotides). The G-rich strand is synthesized by a telomere-specific RT, called telomerase, using a small region of its RNA subunit as template and the 3′-OH on the end of the chromosome as a primer (Blackburn 1992). Found in animals, fungi, and Amoebozoa, TTAGGG was the telomeric simple repeat sequence present in the ancestral Unikont. Moreover, its occurrence in some species of the supergroups Plantae, Chromalveolata, Excavata, and Rhizaria suggests that TTAG GG could be the ancestral telomeric repeat sequence for eukaryotes (Fulnecková et al. 2013).

On the other hand, the use of prokaryotic retroelements to root a RT phylogenetic tree shows that the telomerase seems to have evolved from the RT of an ancestral non-LTR retrotransposon (Eickbush 1997). Furthermore, it is believed that the ability of non-LTR RTs to use the 3′-OH of chromosome ends to prime reverse transcription was crucial for the birth of early telomerases (Moore and Haber 1996; Morrish et al. 2002, 2007; Curcio and Belfort 2007).

Telomerase-based telomeres brought two principal advantages: facilitated telomere homeostasis and a greater structural protection by the incorporation of simple G-rich repeats with the inherent ability to form G-quadruplex structures (Henderson et al. 1987; Arthanari and Bolton 2003; Teixeira and Gilson 2005). G-quadruplexes consist of stacked G-quartets, which are planar arrangements of four guanines held together by Hoogsteen hydrogen bonds (Neidle 2009) (fig. 2*A*). G-quadruplex formation may occur within the terminal G-rich 3′-overhang (fig. 2*B*) or when the overhang invades the adjacent double-stranded region of the telomere to form T-loop structures (fig. 2*C*) (Maizels 2006; Rhodes 2006; Xu et al. 2008; Bochman et al. 2012). Nevertheless, it has been hypothesized that after the appearance of telomerase, the maintenance of telomeres by the primitive T-loop-replication mechanism becomes less relevant (de Lange 2004). The first visualization of telomeric G-quadruplex formation in vivo was performed in the ciliate *Stylonychia* (Schaffitzel et al. 2001; Paeschke et al. 2005). Most recently, a highly specific DNA G-quadruplex antibody has been employed to visualize G-quadruplex structures at human telomeres (Biffi et al. 2013).

It is important to point out that the putative ancestral telomerase-synthesized sequence, $(TTAGGG)_n$, is not only capable of folding into a G-quadruplex structure but is the best one at doing so in vitro (Tran et al. 2011). In addition, recent biophysical studies on the folding of these telomeric G-quadruplexes have shown that structure formation occurs in milliseconds. These folding kinetics are biologically relevant because they are comparable to those of transcription and DNA replication (Zhang and Balasubramanian 2012).

During evolution, mutations in the telomerase RNA template have given rise to repeat variants with different lengths of guanine motifs (G_2, G_4, and more). Recent experiments have found that the G-quadruplexes formed by telomeric repeats with only two consecutive guanines (TTAGG in arthropods and TTAGGC in nematodes) are in equilibrium with G-hairpins and other noncanonical structures (Tran et al. 2011).

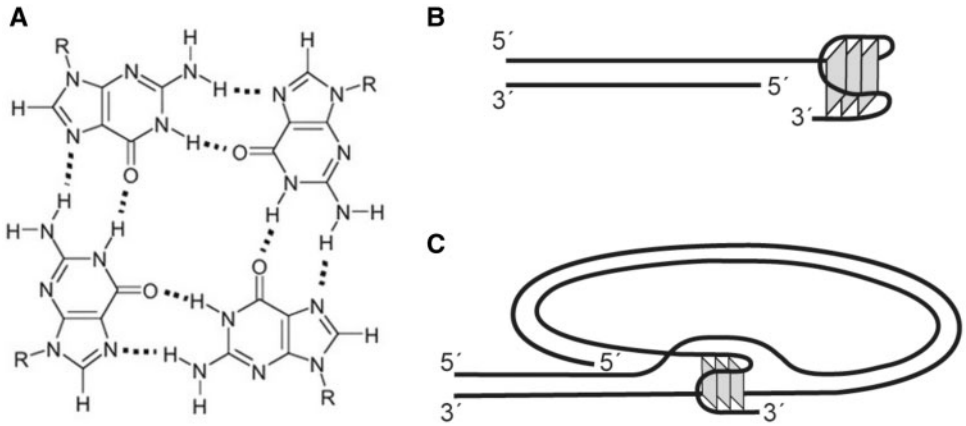


Fig. 2.—Schematic diagrams of a G-quartet and two telomeric G-quadruplexes. (*A*) Four guanines assemble in a planar arrangement to form a G-quartet. Hydrogen bonds are in dashed lines. (*B*) Diagram of an intramolecular G-quadruplex at a telomere end. (*C*) Diagram of a G-quadruplex at a T-loop. The G-quadruplexes in the figure are composed of three stacked G-quartets (shaded squares).

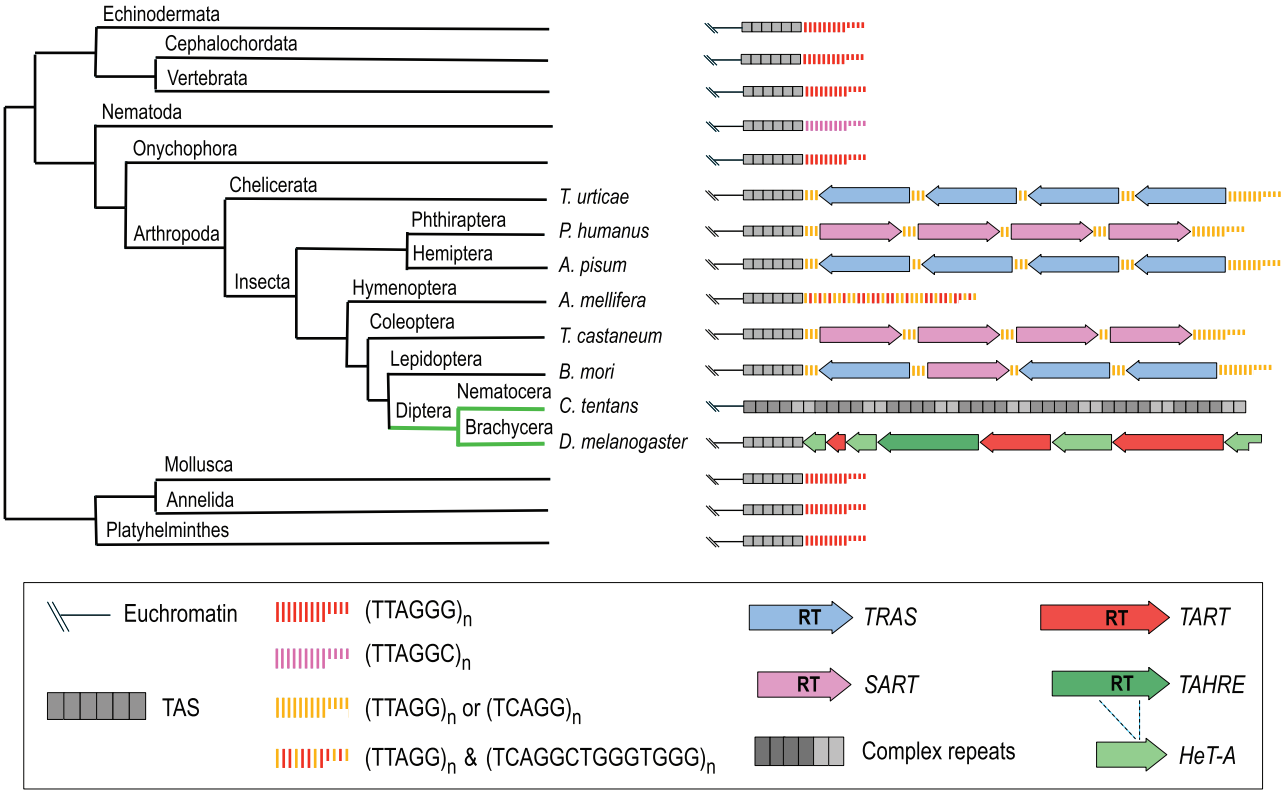


Fig. 3.—Distribution of telomeric sequences within Bilateria. Most eukaryotes have G-rich telomerase-synthesized repeats with adjacent complex subtelomeric repeats called telomere-associated sequences (TAS). In most arthropods, telomere-specific retrotransposons are inserted into telomerase-synthesized repeats. As can be seen in the diagram, *TRAS* elements insert in reverse orientation to that of the *SART* elements. In an ancestor of diptean insects, the telomerase gene was lost (green line). In *Chironomus tentans* (lower Diptera), the telomeric sequences consist of complex tandem repeats maintained by recombination. However, *Drosophila* species have multiple telomere-specific retrotransposons (autonomous and nonautonomous) that transpose to chromosomal ends. The deletion event in the ancestral TAHRE element is shown with dashed lines.

In the silk moth *Bombyx mori* (Lepidoptera) and the flour beetle *Tribolium castaneum* (Coleoptera), the telomerase activity is weak, and telomere-specific non-LTR retroelements (*TRAS* and *SART* family elements in *B. mori* and *SART* family elements in *T. castaneum*) are inserted into the telomeric repeats in a specific manner (Fujiwara et al. 2005; Osanai et al. 2006) that preserves the G/C strand bias (fig. 3). The massive integration of these elements into the proximal regions of the TTAGG repeat arrays of *B. mori* and the TCAGG arrays of *T. castaneum* (an alternative telomere variant in insects) gives rise to huge telomeres with sizes larger than 200 kb.

The telomeres of the honey bee *Apis mellifera* (Hymenoptera) are exceptional among the arthropods because they do not have non-LTR elements inserted into their telomeric repeats. Instead, the telomere sequence consists of TTAGG repeat arrays (Robertson and Gordon 2006) interspersed with TCAGGCTGGG, TCAGGCTGGGTTGGG, and TC AGGCTGGGTGAGGATGGG higher order repeat arrays (Garavís M, Villasante A, unpublished results) (fig. 3). These higher order repeats arose by amplification of the mutated repeats present in proximal telomeric regions and the interspersed pattern developed by further amplifications of the 5-bp repeat arrays together with higher order repeat arrays. However, the TTAGG repeats of *Acyrthosiphon pisum* (Hemiptera) and *Pediculus humanus* (Phthiraptera) contain insertions of non-LTR retrotransposons of the *TRAS* and *SART* family, respectively (International Aphid Genomics Consortium 2010; Kirkness et al. 2010) (fig. 3). As Hemiptera and Phthiraptera are basal to Hymenoptera, Coleoptera, Lepidoptera, and Diptera (fig. 3), the telomeres of *A. mellifera* seem to represent a case where the *TRAS* and/or *SART* retrotransposons were lost at a later stage in evolution. It is tempting to speculate that the appearance of those higher order repeat arrays with propensity to form 3-quartet G-quadruplexes caused the decay, and eventual loss, of the telomeric retrotransposons.

Because the telomeres of the spider mite *Tetranychus urticae* (from the basal branch Chelicerata) are also a mosaic of short TTAGG repeats interrupted by non-LTR retrotransposons closely related to *TRAS* (Grbić et al. 2011) (fig. 3), the telomeres of the arthropods seem to be maintained by telomerase, by insertion of specific non-LTR retrotransposons into the

TTAGG repeat array and by recombination. The same system of telomere maintenance has also been found in some nonarthropod species (Arkhipova and Morrison 2001; Yamamoto et al. 2003; Gladyshev and Arkhipova 2007; Starnes et al. 2012). It has not escaped our notice that the appearance of this apparently suboptimal mechanism of telomere maintenance, which might have created chromosome instability, seems to have coincided with the great arthropod radiation into Chelicerata and Mandibulata.

On the other hand, certain yeasts from the *Ascomycota* phylum have telomeric repeats that are diverse in terms of their sequence, length, and homogeneity (McEachern and Blackburn 1994). In these yeasts, the degenerate repeats result from the nonprocessivity of their telomerases (Prescott and Blackburn 1997). Importantly, these repeats, despite their TG-richness, are less prone to fold into G-quadruplexes (Tran et al. 2011), and it has been shown that in these organisms, the telomere-binding proteins are fast evolving (Teixeira and Gilson 2005). Therefore, it is possible that these yeasts are using an ancestral system of chromosome end protection where the ssDNA-binding proteins facilitate the folding of the 3′-overhangs into G-quadruplex-like structures. This yeast-capping mechanism likely arose de novo by convergent evolution.

Do Noncanonical Secondary Structures Have a Role in the Maintenance of Telomeres without Telomerase?

Once telomerase becomes completely dysfunctional, the gene encoding telomerase could be lost if telomeres are maintained by the ancestral alternative mechanism of homologous recombination. Apparently, this is what happened in the ancestor of Diptera about 260 Ma (Wiegmann et al. 2011). In the lower Diptera, *Anopheles*, *Rhynchosciara,* and *Chironomus*, long tandem repeats are present at chromosome ends, suggesting that telomere maintenance takes place by homologous recombination (Nielsen and Edstrom 1993; Biessmann et al. 1998; Madalena et al. 2010) (fig. 3). In *Drosophila*, however, telomere maintenance occurs primarily by transposition of telomere-specific retrotransposons to receding chromosome ends (fig. 3). In addition to retrotransposition, *Drosophila* telomeres are also maintained, as in any eukaryote, by recombination/gene conversion (Kahn et al. 2000).

In *D. melanogaster*, three telomeric retrotransposons *TART*, *TAHRE,* and *HeT-A* (a nonautonomous element derived from an ancestral *TAHRE* that loss its RT), transpose occasionally to chromosome ends using the free 3′-OH at chromosome termini to prime reverse transcription (Biessmann et al. 1990, 1992; Sheen and Levis 1994; Abad et al. 2004a, 2004b). In agreement with this mechanism, the telomeric elements appear randomly mixed in head-to-tail arrangements and variably truncated at the 5′-end (Mason et al. 2008; Villasante et al. 2008; Pardue and DeBaryshe 2011) (fig. 3). It is noteworthy that deletion of the RT coding region of the telomeric elements has occurred recurrently during *Drosophila* evolution, and multiple nonautonomous elements appear at the telomeres of the *Drosophila* species examined. As an example, up to four nonautonomous elements along with their corresponding autonomous elements have been found in *D. mojavensis* telomeres (Villasante et al. 2007). Interestingly, similar situations occur with group II introns where their RTs also act in *trans* to mobilize multiple deleted introns (Mohr et al. 2010).

Because *Drosophila* telomeres consist of retrotransposon arrays in constant flux, there is not a specific terminal sequence and their telomere-capping proteins (the "terminin" complex) have evolved to bind chromosome ends independently of the primary DNA sequence (Raffa et al. 2009, 2010). The "terminin" complex is functionally analogous to the "shelterin" complex (human telomere-capping proteins), but their components are not evolutionarily conserved (Palm and de Lange 2008; Raffa et al. 2009, 2010). Thus, *Drosophila* telomeres are made of rapidly evolving telomeric retrotransposons (Villasante et al. 2007) and telomere-capping proteins (Gao et al. 2010; Raffa et al. 2010). Moreover, as Verrochio is a telomere-capping protein with one OB-fold domain and all telomeric proteins containing OB folds are 3′-overhang binding proteins, *Drosophila* telomeres also seem to have single-strand overhangs (Raffa et al. 2010).

It is noteworthy that, despite the complexity of telomeric sequences in the genus *Drosophila* and *Chironomus*, the *Drosophila* telomeric retrotransposon arrays and the *Chironomus* telomeric complex repeats also have the telomeric G/C strand bias (Nielsen and Edstrom 1993; Danilevskaya et al. 1998). The conservation of this G/C strand bias may indicate that telomere capping depends on the formation of noncanonical structures based on guanine–guanine interactions. In agreement with this idea, it has been shown that the 3′-untranslated region of the abundant *D. melanogaster* telomeric element *HeT-A* contains sequences with propensity to form G-quadruplexes (Abad and Villasante 1999).

The structural and phylogenetic analyses of all *Drosophila* telomeric-specific retrotransposons show that they had a common ancestor and indicate that non-LTR retrotransposons have been recruited to perform the cellular function of telomere maintenance. Therefore, we propose that the recruitment of *Drosophila* telomeric elements may resemble the ancestral mechanism that led to the maintenance of the "proto-telomeres" of the first eukaryotic chromosomes.

On the other hand, it has been found that yeast cells lacking telomerase can survive telomere sequence loss through the formation of terminal blocks of heterochromatin. This happens by amplifying and rearranging either subtelomeric sequences in *S. cerevisiae* and *S. pombe* or rDNA sequences in *S. pombe* (Lundblad and Blackburn 1993; Jain et al. 2010). Significantly, the *S. cerevisiae* subtelomeric Y′ repeats also have purine/pyrimidine strand bias (Nickles and McEachern 2004), and the *S. pombe* end-protection protein POT1

(protection of telomeres 1) binds, in a nonsequence-specific manner, to the 3′-overhangs of G-rich rDNA (Jain et al. 2010). Interestingly, adaptive recombination-based mechanisms of telomere maintenance (called ALT for alternative lengthening of telomeres) also occur in tumor cells that lack telomerase (Bryan et al. 1995; Cesare and Reddel 2010).

To summarize, in species that have lost telomerase either during evolution (order Diptera) or through experimental manipulation, the data available suggest a role of structural DNA features in telomere maintenance, reveal the importance of telomeric heterochromatin (regardless of the underlying primary sequence) in the recruitment of end-binding proteins, and show how easily backup mechanisms may have been used to maintain telomeres during evolution.

Conclusions

We have discussed how genomes have exploited their noncoding structural potential to establish primordial innovations during eukaryogenesis. In particular, how the highly polymorphic secondary structures based on guanine–guanine interactions have evolved in concert with proteins to allow the origin and evolution of telomeres. In addition, we have hypothesized that the first linear eukaryotic chromosomes arose by the appearance of "proto-telomeres" at DSBs. This ancestral terminal structure had the dual function of end protection and segregation. Furthermore, we have discussed that the concomitant "proto-telomere" heterochromatin formation was fundamental for this key evolutionary innovation. Once again, the study of a noncanonical mechanism, like the maintenance of *Drosophila* telomeres, has generated new insights into the evolution of eukaryotes.

Acknowledgments

The authors thank Jim Mason for discussions on the origin of *Drosophila* telomeres, Maria J. Puertas for discussions on the heterochromatin of meiotic neocentromeres, and Douglas V. Laurents for revision of the manuscript. They are grateful to the reviewers for their valuable comments and suggestions. They apologize to those whose work was not cited in this review. This work was supported by the FPI fellowship BES-2009-027909 from the Ministerio de Ciencia e Innovación to M.G., by the Ministerio de Ciencia e Innovación (CTQ2010-21567-C02-02) to C.G., the Ministerio de Economía y Competitividad (BFU2011-30295-C02-01) to A.V., and by an institutional grant from the Fundación Ramón Areces to the Centro de Biología Molecular "Severo Ochoa."

Literature Cited

Abad JP, et al. 2004a. Genomic analysis of *Drosophila melanogaster* telomeres: full-length copies of HeT-A and TART elements at telomeres. Mol Biol Evol. 21:1613–1619.

Abad JP, et al. 2004b. TAHRE, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. Mol Biol Evol. 21:1620–1624.

Abad JP, Villasante A. 1999. The 3′ non-coding region of the *Drosophila melanogaster* HeT-A telomeric retrotransposon contains sequences with propensity to form G-quadruplex DNA. FEBS Lett. 453:59–62.

Agudo M, et al. 1999. Centromeres from telomeres? The centromeric region of the Y chromosome of *Drosophila melanogaster* contains a tandem array of telomeric HeT-A- and TART-related sequences. Nucleic Acids Res. 27:3318–3324.

Arkhipova IR, Morrison HG. 2001. Three retrotransposon families in the genome of *Giardialamblia*: two telomeric, one dead. Proc Natl Acad Sci U S A. 98:14497–14502.

Arthanari H, Bolton PH. 2003. Did quadruplex DNA play a role in the evolution of the eukaryotic linear chromosome? Mini Rev Med Chem. 3:1–9.

Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J. 2007. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. Science 318:798–780.

Biessmann H, et al. 1990. Addition of telomere-associated HeT DNA sequences "heals" broken chromosome ends in *Drosophila*. Cell 61: 663–763.

Biessmann H, et al. 1992. HeTA, a transposable element specifically involved in "healing" broken chromosome ends in *Drosophila melanogaster*. Mol Cell Biol. 12:3910–3918.

Biessmann H, Kobeski F, Walter MF, Kasravi A, Roth CW. 1998. DNA organization and length polymorphism at the 2L telomeric region of *Anopheles gambiae*. Insect Mol Biol. 7:83–93.

Biffi G, Tannahill D, McCafferty J, Balasubramanian S. 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. Nat Chem. 5:182–186.

Blackburn EH. 1992. Telomerases. Annu Rev Biochem. 61:113–129.

Blocker FJ, et al. 2005. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. RNA 11: 14–28.

Bochman ML, Paeschke K, Zakian VA. 2012. DNA secondary structures: stability and function of G-quadruplex structures. Nat Rev Genet. 13: 770–780.

Bryan TM, Englezou A, Gupta J, Bacchetti S, Reddel RR. 1995. Telomere elongation in immortal human cells without detectable telomerase activity. EMBO J. 4:4240–4248.

Cavalier-Smith T. 1981. The origin and early evolution of the eukaryotic cell. In: Carlile MJ, Collins JF, Moseley BEB, editors. Molecular and cellular aspects of microbial evolution. Cambridge: Cambridge University Press. p. 33–84.

Cavalier-Smith T. 2010. Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution. Biol Direct. 5:7.

Cesare AJ, Reddel RR. 2010. Alternative lengthening of telomeres: models, mechanisms and implications. Nat Rev Genet. 1:319–330.

Chen G, Rubinstein B, Li R. 2012. Whole chromosome aneuploidy: big mutations drive adaptation by phenotypic leap. Bioessays 34: 893–900.

Choi ES, et al. 2012. Factors that promote H3 chromatin integrity during transcription prevent promiscuous deposition of CENP-A (Cnp1) in fission yeast. PLoS Genet. 8(9):e1002985.

Curcio MJ, Belfort M. 2007. The beginning of the end: links between ancient retroelements and modern telomerases. Proc Natl Acad Sci U S A. 104:9107–9108.

Danilevskaya ON, Lowenhaupt K, Pardue ML. 1998. Conserved subfamilies of the *Drosophila* HeT-A telomere-specific retrotransposon. Genetics 148:233–242.

Deng Z, Norseen J, Wiedmer A, Riethman H, Lieberman PM. 2009. TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. Mol Cell. 35:403–413.

Eickbush TH. 1997. Telomerase and retrotransposons: which came first? Science 277:911–912.
Flynn RL, et al. 2011. TERRA and hnRNPA1 orchestrate an RPA-to-POT1 switch on telomeric single-stranded DNA. Nature 471:532–536.
Fujiwara H, Osanai M, Matsumoto T, Kojima KK. 2005. Telomere specific non-LTR retrotransposons and telomere maintenance in the silkworm, *Bombyx mori*. Chromosome Res. 13:455–467.
Fulnecková J, et al. 2013. A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. Genome Biol Evol. 5: 468–483.
Gao G, et al. 2010. HipHop interacts with HOAP and HP1 to protect *Drosophila* telomeres in a sequence-independent manner. EMBO J. 29:819–829.
Gelinas AD, et al. 2009. Telomere capping proteins are structurally related to RPA with an additional telomere-specific domain. Proc Natl Acad Sci U S A. 106:19298–19303.
Gladyshev EA, Arkhipova IR. 2007. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. Proc Natl Acad Sci U S A. 104:9352–9357.
Grbić M, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 479:487–492.
Heddi B, Phan AT. 2011. Structure of human telomeric DNA in crowded solution. J Am Chem Soc. 133:9824–9833.
Henderson E, Hardin CC, Walk SK, Tinoco I Jr, Blackburn EH. 1987. Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs. Cell 51:899–908.
Heun P, et al. 2006. Mislocalization of the *Drosophila* centromere-specific histone CID promotes formation of functional ectopic kinetochores. Dev Cell. 10:303–315.
International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol. 8:e1000313.
Ishii K, et al. 2008. Heterochromatin integrity affects chromosome reorganization after centromere dysfunction. Science 321:1088–1091.
Jain D, Hebden AK, Nakamura TM, Miller KM, Cooper JP. 2010. HAATI survivors replace canonical telomeres with blocks of generic heterochromatin. Nature 467:223–227.
Kahn T, Savitsky M, Georgiev P. 2000. Attachment of HeT-A sequences to chromosomal termini in *Drosophila melanogaster* may occur by different mechanisms. Mol Cell Biol. 20:7634–7642.
Kirkness EF, et al. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A. 107:12168–12173.
Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol Direct. 1:22.
Koonin EV. 2011. The logic of chance: the nature and origin of biological evolution. Upper Saddle River (NJ): FT Press.
Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. Annu Rev Genet. 38:1–35.
de Lange T. 2004. T-loops and the origin of telomeres. Nat Rev Mol Cell Biol. 5:323–329.
Lannan FM, Mamajanov I, Hud NV. 2012. Human telomere sequence DNA in water-free and high-viscosity solvents: G-quadruplex folding governed by Kramers rate theory. J Am Chem Soc. 134:15324–15330.
Lefrançois P, Auerbach RK, Yellman CM, Roeder GS, Snyder M. 2013. Centromere-like regions in the budding yeast genome. PLoS Genet. 9:e1003209.
Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72:595–605.
Lundblad V, Blackburn EH. 1993. An alternative pathway for yeast telomere maintenance rescues est1-senescence. Cell 73:347–360.
Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404.
Madalena CRG, Amabis JM, Gorab E. 2010. Unusually short tandem repeats appear to reach chromosome ends of *Rhynchosciara americana* (Diptera: Sciaridae). Chromosoma 119:613–623.
Maizels N. 2006. Dynamic roles for G4 DNA in the biology of eukaryotic cells. Nat Struct Mol Biol. 13:1055–1059.
Marshall OJ, Chueh AC, Wong LH, Choo KH. 2008. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. Am J Hum Genet. 82:261–282.
Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. Nature 440:41–45.
Mason JM, Frydrychova RC, Biessmann H. 2008. *Drosophila* telomeres: an exception providing new insights. Bioessays 30:25–37.
McClintock B. 1984. The significance of responses of the genome to challenge. Science 226:792–801.
McEachern MJ, Blackburn EH. 1994. A conserved sequence motif within the exceptionally diverse telomeric sequences of budding yeasts. Proc Natl Acad Sci U S A. 91:3453–3457.
Méndez-Lago M, et al. 2009. Novel sequencing strategy for repetitive DNA in a *Drosophila* BAC clone reveals that the centromeric region of the Y chromosome evolved from a telomere. Nucleic Acids Res. 37: 2264–2273.
Mohr G, Ghanem E, Lambowitz AM. 2010. Mechanisms used for genomic proliferationby thermophilic group II introns. PLoS Biol. 8(6):e1000391.
Moore JK, Haber JE. 1996. Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature 383:644–646.
Morrish TA, et al. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet. 2:159–165.
Morrish TA, et al. 2007. Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature 446:208–212.
Murphy WJ, et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309:613–617.
Neidle S. 2009. The structures of quadruplex nucleic acids and their complexes. Curr Opin Struct Biol. 9:239–250.
Nickles K, McEachern MJ. 2004. Characterization of *Kluyveromyces lactis* subtelomeric sequences including a distal element with strong purine/pyrimidine strand bias. Yeast 21:813–830.
Nielsen L, Edstrom JE. 1993. Complex telomere-associated repeat units in members of the genus *Chironomus* evolve from sequences similar to simple telomeric repeats. Mol Cell Biol. 13:1583–1589.
Olszak AM, et al. 2011. Heterochromatin boundaries are hotspots for *de novo* kinetochore formation. Nat Cell Biol. 13:799–808.
Osanai M, Kojima KK, Futahashi R, Yaguchi S, Fujiwara H. 2006. Identification and characterization of the telomerase reverse transcriptase of *Bombyx mori* (silkworm) and *Tribolium castaneum* (flour beetle). Gene 376:281–289.
Paeschke K, Simonsson T, Postberg J, Rhodes D, Lipps HJ. 2005. Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. Nat Struct Mol Biol. 12:847–854.
Palm W, de Lange T. 2008. How shelterin protects mammalian telomeres. Annu Rev Genet. 42:301–334.
Pardue ML, DeBaryshe PG. 2011. Retrotransposons that maintain chromosome ends. Proc Natl Acad Sci. U S A. 108:20317–20324.
Prescott J, Blackburn EH. 1997. Telomerase RNA mutations in *Saccharomyces cerevisiae* alter telomerase action and reveal nonprocessivity *in vivo* and *in vitro*. Genes Dev. 11:528–540.
Puertas MJ, Villasante A. 2013. Is the heterochromatin of meiotic neocentromeres a remnant of the early evolution of the primitive centromere?. In: Jiang J, Birchler JA, editors. Plant centromere biology Oxford (UK): Wiley-Blackwell. p. 95–109.
Raffa GD, et al. 2009. The *Drosophila* modigliani (moi) gene encodes a HOAP-interacting protein required for telomere protection. Proc Natl Acad Sci U S A. 106:2271–2276.

Genome Biology and Evolution SMBE

Raffa GD, et al. 2010. Verrocchio, a *Drosophila* OB fold-containing protein, is a component of the terminin telomere-capping complex. Genes Dev. 24:1596–1601.

Rhodes D. 2006. The structural biology of telomeres. In: de Lange T, Lundblad V, Blackburn EH, editors. Telomeres, 2nd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. p. 317–343.

Robertson HM, Gordon KH. 2006. Canonical TTAGG-repeat telomeres and telomerase in the honey bee, *Apis mellifera*. Genome Res. 16: 1345–1351.

Schaffitzel C, et al. 2001. In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with *Stylonychia lemnae* macronuclei. Proc Natl Acad Sci U S A. 98:8572–8577.

Schoeftner S, Blasco MA. 2008. Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat Cell Biol. 10:228–236.

Sharp PA. 1985. On the origin of RNA splicing and introns. Cell 42: 397–400.

Sharp PA. 1991. "Five easy pieces". Science 254:663.

Sheen FM, Levis RW. 1994. Transposition of the LINE-like retrotransposon TART to *Drosophila* chromosome termini. Proc Natl Acad Sci U S A. 91: 12510–12514.

Shpiz S, et al. 2011. Mechanism of the piRNA-mediated silencing of *Drosophila* telomeric retrotransposons. Nucleic Acids Res. 39: 8703–8711.

Starnes JH, Thornbury DW, Novikova OS, Rehmeyer CJ, Farman ML. 2012. Telomere-targeted retrotransposons in the rice blast fungus *Magnaporthe oryzae*: agents of telomere instability. Genetics 2: 389–406.

Teixeira MT, Gilson E. 2005. Telomere maintenance, function and evolution: the yeast paradigm. Chromosome Res. 13:535–548.

Tran PL, Mergny JL, Alberti P. 2011. Stability of telomeric G-quadruplexes. Nucleic Acids Res. 39:3282–3294.

Ventura M, et al. 2003. Neocentromeres in 15q24–26 map to duplicons which flanked an ancestral centromere in 15q25. Genome Res. 13: 2059–2068.

Ventura M, et al. 2004. Recurrent sites for new centromere seeding. Genome Res. 14:1696–1703.

Villasante A, Abad JP, Méndez-Lago M. 2007. Centromeres were derived from telomeres during the evolution of the eukaryotic chromosome. Proc Natl Acad Sci U S A. 104:10542–10547.

Villasante A, de Pablos B, Mendez-Lago M, Abad JP. 2008. Telomere maintenance in *Drosophila*: rapid transposon evolution at chromosome ends. Cell Cycle 7:2134–2138.

Villasante A, et al. 2007. *Drosophila* telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. Genome Res. 17:1909–1918.

Villasante A, Méndez-Lago M., Abad JP, Montejo de Garcini E. 2007. The birth of the centromere. Cell Cycle 6:2872–2876.

Wiegmann BM, et al. 2011. Episodic radiations in the fly tree of life. Proc Natl Acad Sci U S A. 108:5690–5695.

Xiong Y, Eickbush T. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 9: 3353–3362.

Xu L, Feng S, Zhou X. 2011. Human telomeric G-quadruplexes undergo dynamic conversion in a molecular crowding environment. Chem Commun (Camb). 47:3517–3519.

Xu Y, Sato H, Sannohe Y, Shinohara K, Sugiyama H. 2008. Stable lariat formation based on a G-quadruplex scaffold. J Am Chem Soc. 130: 6470–16471.

Yamamoto Y, et al. 2003. Retrotransposon-mediated restoration of *Chlorella* telomeres: accumulation of Zepp retrotransposons at termini of newly formed minichromosomes. Nucleic Acids Res. 31: 4646–4653.

Yona AH, et al. 2012. Chromosomal duplication is a transient evolutionary solution to stress. Proc Natl Acad Sci U S A. 109: 21010–21015.

Zeitlin SG, et al. 2009. Double-strand DNA breaks recruit the centromeric histone CENP-A. Proc Natl Acad Sci U S A. 106:15762–15767.

Zhang AY, Balasubramanian S. 2012. The kinetics and folding pathways of intramolecular g-quadruplex nucleic acids. J Am Chem Soc. 134: 19297–19308.

Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995. Group II intron mobility occurs by target DNA-primed reverse transcription. Cell 82: 545–554.

Genome Biology and Evolution

SMBE

Artículo 2.

Mechanical unfolding of long human telomeric RNA (TERRA)

Miguel Garavís, Rebeca Bocanegra, Elías Herrero-Galán, Carlos González, Alfredo Villasante and J. Ricardo Arias-González

Desplegamiento mecánico de RNA telomérico humano (TERRA)

Miguel Garavís, Rebeca Bocanegra, Elías Herrero-Galán, Carlos González, Alfredo Villasante and J. Ricardo Arias-González

El RNA telomérico (TERRA) es un RNA no codificante que forma parte de la heterocromatina telomérica. TERRA tiene una extensión variable presentando una media de 34 repeticiones de la secuencia telomérica r(GGGUUA). Diversos estudios estructurales han mostrado que secuencias cortas formadas por varias repeticiones de RNA telomérico son capaces de plegarse formando estructuras tipo cuádruplex de guanina *in vitro*. En este trabajo presentamos el estudio de una secuencia de RNA telomérico formada por 16 repeticiones del motivo de repetición del telómero humano r(GGGUUA). El diseño experimental descrito en este capítulo permitió el estudio de esta molécula mediante técnicas biofísicas, como RMN y dicroísmo circular, así como mediante una técnica de molécula única como las pinzas ópticas.

La síntesis y purificación a gran escala de un TERRA formado por 16 repeticiones hizo posible la obtención de espectros de ^{1}H-RMN de dicha molécula que revelaron que ésta se pliega formando estructuras tipo cuádruplex de guanina y que, probablemente, existan interacciones entre los lazos o *loops* que separan los sucesivos cuádruplex en la estructura. Los resultados de dicroísmo circular indican que la molécula se pliega formando cuatro cuádruplex de conformación paralela.

Los estudios mediante pinzas ópticas permitieron observar la respuesta a la extensión de la secuencia formada por 16 repeticiones r(GGGUUA) y de otras dos secuencias formadas por 5 y por 25 repeticiones. El análisis de las curvas de fuerza-extensión revela que la fuerza necesaria para la apertura de un cuádruplex de guanina de RNA es menor que la que provoca la apertura de un cuádruplex de guanina de DNA. Así mismo, en las secuencias de 16 y 25 repeticiones se observaron eventos de apertura cooperativa lo que indica que en dichas moléculas existe una interacción entre unidades de cuádruplex consecutivas.

Aportación personal al trabajo: Preparación de los RNAs teloméricos para el estudio por RMN, CD y pinzas ópticas. Realización e interpretación de los experimentos de RMN y CD de las secuencias de TERRA. Escritura y discusión del manuscrito.

ChemComm

RSC Publishing

COMMUNICATION

View Article Online
View Journal | View Issue

Mechanical unfolding of long human telomeric RNA (TERRA)†

Cite this: *Chem. Commun.*, 2013, **49**, 6397

Received 22nd April 2013,
Accepted 28th May 2013

DOI: 10.1039/c3cc42981d

www.rsc.org/chemcomm

Miguel Garavís,‡[ab] Rebeca Bocanegra,‡[c] Elías Herrero-Galán,[c] Carlos González,*[b] Alfredo Villasante*[a] and J. Ricardo Arias-Gonzalez*[de]

We report the first single molecule investigation of TERRA molecules. By using optical-tweezers and other biophysical techniques, we have found that long RNA constructions of up to 25 GGGUUA repeats form higher order structures comprised of single parallel G-quadruplex blocks, which unfold at lower forces than their DNA counterparts.

Telomeres are nucleoprotein structures that protect chromosome ends from being recognized as DNA breaks.[1] Mammalian telomere DNA consists of long tandem arrays of double-stranded TTAGGG repeats that are maintained by the telomerase.[2] The 3′ end of each telomere terminates in a G-rich single-stranded overhang (150–200 nucleotides) that is able to self-fold into G-quadruplexes. The demonstration of the *in vivo* presence of DNA G-quadruplexes at telomeres suggests that these non-canonical secondary structures have a role in telomere end-protection, and therefore in chromosome stability.[3,4] It has also been found that telomere end-binding proteins control the folding and unfolding of DNA G-quadruplexes *in vitro*[5] and *in vivo*.[4]

Telomeres are transcribed from the subtelomeric regions towards chromosome ends into telomeric repeat-containing RNA (TERRA).[6,7] These non-coding RNA molecules contain subtelomere-derived sequences and an average of 34 GGGUUA repeats at their 3′ end.[8] TERRA acts as a scaffold for the assembly of telomeric proteins involved in telomere maintenance and telomeric heterochromatin formation.[9,10] Importantly, there is also evidence for the presence of TERRA RNA G-quadruplexes in living cells.[11] It has been shown by nuclear magnetic resonance (NMR) spectroscopy[12,13] and by X-ray crystallography[14] that short TERRA molecules form parallel-stranded G-quadruplex structures but high-resolution structures of long telomeric RNA have not yet been obtained. Electrospray mass spectrometry[15] and computational analysis[12,16] of long TERRA molecules have suggested stacking between their quadruplexes. The largest TERRA molecule (51nt) analyzed by NMR has shown intramolecular parallel G-quadruplexes.[16]

Here, we report the synthesis and purification of TERRA molecules up to 96nt (twice the largest molecule previously analyzed by NMR), a size similar to the endogenous nuclear TERRA[8] (see ESI†). We have also used optical-tweezers (OT) to study RNA molecules with different numbers of possible quadruplexes ranging from 1 (Q1) to 6 (Q6).

CD spectra of the RNA sequence Q4 in 25 mM potassium phosphate pH 7 buffer showed a positive band at 260 nm and a negative band at 240 nm (Fig. 1a), the characteristic signature of parallel-stranded G-quadruplex conformations. This result is in agreement with previously reported observations showing that other telomeric RNA sequences are folded into parallel G-quadruplex structures both in sodium and potassium solutions.[12,13,16–20] It is also important to mention that no band at 300 nm was observed for this long telomeric RNA as it was reported for 22-mer and 45-mer telomeric RNA in ammonium acetate buffer.[15] On the other hand, the normalized CD spectra of the bimolecular G-quadruplex formed by r(UAGGGUUAGGGU) and Q4 showed a similar shape and amplitude, which indicates that most of the Q4 molecules form four G-quadruplexes (Fig. 1b).[21]

Thermal stability of Q4 was also measured by recording CD melting curves. To test the influence of the potassium concentration on the thermal stability we performed the experiments in two different potassium-containing buffers (Fig. 1c). Melting temperatures of 65.5 °C and 72.5 °C were obtained in 25 mM potassium phosphate pH 7 and 25 mM potassium phosphate pH 7 buffer plus 50 mM KCl respectively. Such a notable difference between melting temperatures confirms the influence of potassium concentration on the thermal stability of RNA G-quadruplexes.[22] On the other hand, comparison of melting temperatures of Q4 and intramolecular telomeric RNA sequences reported in the

[a] *Instituto de Química Física Rocasolano, CSIC, C/Serrano 119, 28006 Madrid, Spain. E-mail: cgonzalez@iqfr.csic.es; Fax: +34 915642431; Tel: +34 915619400*
[b] *Centro de Biología Molecular "Severo Ochoa" CSIC-UAM, Universidad Autónoma de Madrid, Nicolás Cabrera 1, 28049 Madrid, Spain. E-mail: avillasante@cbm.uam.es; Fax: +34 911964420; Tel: +34 911964401*
[c] *Centro Nacional de Biotecnología (CNB-CSIC), C/Darwin 3, Cantoblanco, 28049 Madrid, Spain*
[d] *IMDEA Nanociencia, C/Faraday 9, Cantoblanco, 28049 Madrid, Spain. E-mail: ricardo.arias@imdea.org; Fax: +34 912998725; Tel: +34 912998860*
[e] *CNB-CSIC-IMDEA Nanociencia Associated Unit "Unidad de Nanobiotecnología", Spain*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3cc42981d

‡ M. Garavís and R. Bocanegra made equal contribution to this paper.

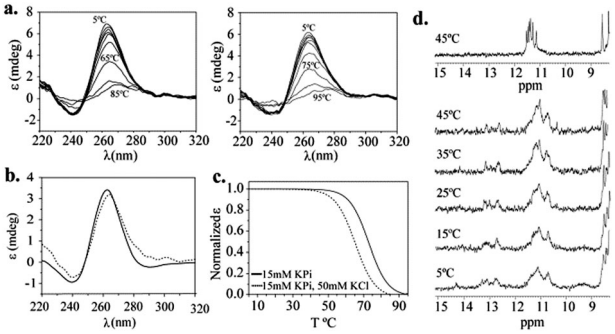


Fig. 1 (a) Series of CD spectra of Q4 in 15 mM KPi pH 7 buffer (left) and 15 mM KPi pH 7, 50 mM KCl (right) at different temperatures, [oligonucleotide] = 2 μM. (b) CD melting curves of Q4 at different K^+ concentrations. (c) CD spectra normalized for G-quadruplex concentration of r(UAGGGUUAGGGU) (straight line) and Q4 (dotted line) at 25 °C in buffer 25 mM KPi pH 7 and 15 mM KPi pH 7 respectively. (d) NMR spectra of r(UAGGGUUAGGGU) at 45 °C (top) and Q3 (15 GGGUUA repeats) at different temperatures (bottom) in H_2O/D_2O 9 : 1 in 15 mM KPi pH 7 buffer. RNA concentration of 100 μM and 10 μM, respectively.

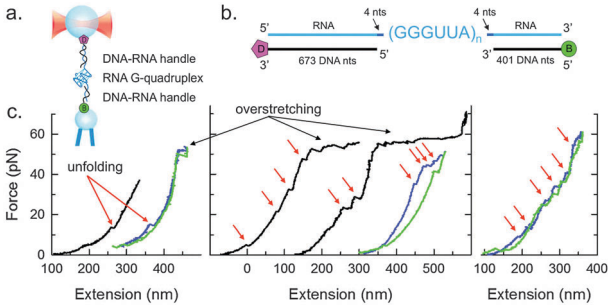


Fig. 2 (a) Cartoon of the experiment: a Q_n construction is attached between a bead (biotin–streptavidin linkage), held by suction on the top of a micropipette, and an optically trapped bead (digoxigenin–anti-digoxigenin antibody linkage).[28] (b) Cartoon of the OT construction showing each component and its length; biotin (B) and digoxigenin (D). (c) Force–extension curves showing unfolding events (red arrows) of Q1 (left), Q4 (middle) and Q6 (right) preparations. Black and blue traces show stretching assays and green traces are relaxation paths corresponding to the blue traces. Each stretch–relaxation curve was offset ~100 nm to increase the clarity of the figure. Black arrows mark the beginning of the overstretching transition.

bibliography[17,18] suggests that the length of the sequence is not necessarily a factor that enhances thermal stability.

Large scale synthesis and purification of long telomeric RNA were performed following previously described protocols[23] yielding enough amount of sample to acquire 1D H-NMR spectra. Despite the higher widening of the signals produced by the size of the molecule, the shape of the NMR spectra is comparable to the one observed for r(UAGGGUUAGGGU), showing signals between 10 and 12 ppm, the characteristic chemical shifts of guanine imino protons forming G-tetrads (Fig. 1d). Moreover, the imino proton signals remain visible at 55 °C indicating a remarkable thermostability, consistent with G-quadruplex folding. Interestingly, imino proton signals at around 13 ppm are also visible in the whole range of temperature indicating contacts involving nucleotides of the loops. To our knowledge, the presence of these signals has not been observed in other NMR studies on telomeric RNA sequences. This fact suggests the occurrence of loop–loop interactions between different quadruplexes. These interactions are consistent with A:U/T base pairs observed in the X-ray structures of the telomeric RNA[14] or DNA quadruplex.[24]

The vectors employed to produce telomeric RNA in sufficiently large amounts for NMR experiments were also used to generate the constructs required for OT analysis (Fig. 2a and ESI†). Thus, three constructions with 5 (Q1), 16 (Q4) and 25 (Q6) telomeric repeats were obtained. The telomeric sequences were sandwiched between two hybrid duplex RNA:DNA acting as handles (Fig. 2b). Single molecules were stretched with OT to forces exceeding the overstretching transition of the hybrid handles,[25,26] following the scheme of Fig. 2a.

Fig. 2c shows, from left to right, three series of single-molecule force–extension experiments corresponding to Q1, Q4 and Q6 sample preparations, respectively. In the left panel, two experiments, each with a different molecule, are shown. Single-unfolding events are observed in accord with what was described for DNA quadruplexes.[27–29] Typical rupture forces are 15 ± 2 pN, below those observed for DNA quadruplexes for the experimental loading rate used here (22 ± 2 pN s^{-1}, see ESI†).[27,30] Weaker stacking interactions between adjacent G-tetrad planes in RNA quadruplexes with respect to those of DNA may be responsible for the lower force needed to unfold TERRA with respect to comparable DNA quadruplex motifs. A similar scenario was found for the base-stacking interactions in double-stranded (ds) nucleic acids, where the dsRNA molecules showed a lower stretch modulus in the force–extension curves with respect to their sequence-equivalent DNA counterparts.[25] It is also important to note that TERRA involves only parallel arrangements, which for DNA quadruplexes have been described with lower unfolding forces than anti-parallel ones.[27] The stretch–relaxation (blue-green) experimental curve of Fig. 2c also shows that unfolding and refolding processes in RNA G-quadruplexes are not reversible, as reflected by the fact that relaxation traces did not superimpose over the stretching paths (blue and green curves in Fig. 2c). A similar irreversibility was also previously exhibited by DNA G-quadruplexes.[27]

Fig. 2c, middle panel, shows three experiments in which sequential and simultaneous rupture events take place. The leftmost curve of this panel shows four sequential unfolding events, the last one taking place right before the overstretching transition of the hybrid handles, below 60 pN. The middle curve shows two unfolding events whose extension coordinate exceeds the mean unfolding distances (see below), probably due to the simultaneous opening of two G-quadruplexes.[31] The rightmost, blue-green experimental traces in Fig. 2c (middle) correspond to a stretch–relaxation cycle in which sequential unfolding events are almost simultaneous. The rightmost and leftmost curves of this middle panel correspond to stretch–relaxation cycles of the same molecule. This sample molecule shows additional features of the mechanical behaviour of the G-quadruplexes, namely, that G-quadruplexes do refold upon relaxation but that rupture events during a new stretching cycle do not necessarily overlap with those from previous cycles. Finally, Fig. 2c, right panel, shows a stretch–relaxation cycle of a Q6 single-molecule sample. In this case, six sequential unfolding events and their corresponding refolding ones are visible in the curves. The stretch–relaxation cycle shows, as before, a hysteretic behaviour.

RNA preparations are labile, which typically resulted in molecular breakage of the single-molecule constructs at high forces. However, some of the constructions could resist forces beyond the overstretching transition of the hybrid handles.

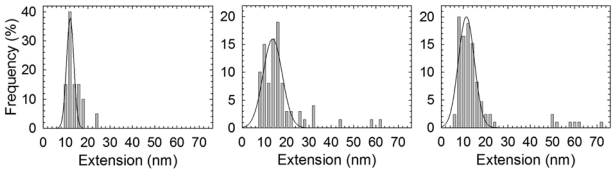


Fig. 3 Unfolding extension histograms for Q1 (left, n = 20), Q4 (center, n = 68) and Q6 (right, n = 85) TERRA repeats from single-molecule experiments. Peak position (and standard deviation) according to Gaussian fits to the experimental histograms are 12.1 (1.6) nm, 13.7 (4.2) nm and 11.3 (3.6) nm, respectively.

This transition provides here a control check to show that single-molecule constructs were used in the OT experiments. Analysis of this transition shows that hybrid DNA:RNA molecules approximately overstretch as A-form molecules.[25]

The fact that the TERRA G-quadruplex showed a clear parallel conformation is reflected in the distribution of extension change upon unfolding. Fig. 3, left panel, shows that the change in extension for Q1 preparation peaks at ~12 nm, with no apparent mixture of populations.[27,31] This distance is compatible with the unfolding of one quadruplex and the presence of 14 nucleotides, distributed as one GGGUUA repeat plus 4 flanking nucleotides at both ends (see Fig. 2b and ESI†), which may interact with the RNA G-quadruplex at low force in Q1 preparation.

When several tandem quadruplexes are unfolded, the extension distribution peaks at approximately the same distance (Fig. 3, middle and right panels) but the standard deviation is larger. The conserved mean unfolding distance shows that there is a preference for sequential unfolding and the increase in standard deviation reflects that interaction between quadruplexes takes place when they are arranged in tandem repeats, as discussed above. In agreement with former literature on DNA quadruplexes,[32] cooperative unfolding events took place in our experiments (see Fig. 3 *center* and *right*) but were infrequent compared to sequential ones for TERRA tandem repeats. The unfolding sequence of rupture events in the force–extension traces is thus more similar to those shown for protein unfolding in tandem preparations[33] than other cooperative transitions in DNA or RNA (overstretching or unzipping).[26]

TERRA is a promising cancer therapeutic target because it is required for telomeric heterochromatin formation and because its higher-order G-quadruplex structure provides specific binding sites to optimize drug design. The single molecule characterization of TERRA molecules established here provides useful mechanistic information that can be readily used to find specific TERRA–ligand interactions, as well as in future studies on TERRA–proteins and TERRA–DNA interactions.

In conclusion, mechanical unfolding experiments of long RNA telomeric repeats (TERRA) indicate that these molecules form homogeneous parallel quadruplexes with typical rupture forces lower than their DNA counterparts. We have found that TERRA forms higher order structures at a single molecule level, which exhibit some cooperative unfolding events.

This work was supported by grants from the Spanish Ministry of Science and Innovation (grants RYC2007-01765 to JRA-G, BFU2011-30295-C02-01 to AV, and CTQ2010-21567-C02-02 to CG). MG was supported by the FPI fellowship BES-2009-027909. RB and EH-G were supported by Comunidad de Madrid, grant CAM-S2009MAT-1507. AV acknowledges an institutional grant from the Fundación Ramón Areces to the CBMSO. JRA-G wants to thank Prof. J. L. Carrascosa and Prof. J. M. Valpuesta (CNB-CSIC) for their continuous support and encouragement in this research. We also acknowledge the excellent technical assistance of Beatriz de Pablos (CBMSO).

Notes and references

1 T. de Lange, *Genes Dev.*, 2005, **19**, 2100.
2 E. H. Blackburn, *Nature*, 1991, **350**, 569.
3 G. Biffi, D. Tannahill, J. McCafferty and S. Balasubramanian, *Nat. Chem.*, 2013, **5**, 182.
4 K. Paeschke, T. Simonsson, J. Postberg, D. Rhodes and H. J. Lipps, *Nat. Struct. Mol. Biol.*, 2005, **12**, 847.
5 H. Hwang, N. Buncher, P. L. Opresko and S. Myong, *Structure*, 2012, **20**, 1872.
6 C. M. Azzalin, P. Reichenbach, L. Khoriauli, E. Giulotto and J. Lingner, *Science*, 2007, **318**, 798.
7 S. Schoeftner and M. A. Blasco, *Nat. Cell Biol.*, 2008, **10**, 228.
8 A. Porro, S. Feuerhahn, P. Reichenbach and J. Lingner, *Mol. Cell. Biol.*, 2010, **30**, 4808.
9 Z. Deng, J. Norseen, A. Wiedmer, H. Riethman and P. M. Lieberman, *Mol. Cell*, 2009, **35**, 403.
10 I. L. de Silanes, M. S. d'Alcontres and M. A. Blasco, *Nat. Commun.*, 2010, **1**, 33.
11 Y. Xu, Y. Suzuki, K. Ito and M. Komiyama, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 14579.
12 H. Martadinata and A. T. Phan, *J. Am. Chem. Soc.*, 2009, **131**, 2570.
13 Y. Xu, K. Kaminaga and M. Komiyama, *J. Am. Chem. Soc.*, 2008, **130**, 11179.
14 G. W. Collie, S. M. Haider, S. Neidle and G. N. Parkinson, *Nucleic Acids Res.*, 2010, **38**, 5569.
15 G. W. Collie, G. N. Parkinson, S. Neidle, F. Rosu, E. De Pauw and V. Gabelica, *J. Am. Chem. Soc.*, 2010, **132**, 9328.
16 H. Martadinata, B. Heddi, K. W. Lim and A. T. Phan, *Biochemistry*, 2011, **50**, 6455.
17 A. Arora and S. Maiti, *J. Phys. Chem. B*, 2009, **113**, 10515.
18 A. Joachimi, A. Benz and J. S. Hartig, *Bioorg. Med. Chem.*, 2009, **17**, 6811.
19 J. Qi and R. H. Shafer, *Biochemistry*, 2007, **46**, 7599.
20 A. Randall and J. D. Griffith, *J. Biol. Chem.*, 2009, **284**, 13980.
21 L. Petraccone, in *Quadruplex Nucleic Acids*, ed. J. B. Chaires and D. Graves, Springer, Berlin, Heidelberg, 2013, vol. 330, p. 23.
22 S. Kumari, A. Bugaut and S. Balasubramanian, *Biochemistry*, 2008, **47**, 12664.
23 S. A. McKenna, I. Kim, E. V. Puglisi, D. A. Lindhout, C. E. Aitken, R. A. Marshall and J. D. Puglisi, *Nat. Protoc.*, 2007, **2**, 3270.
24 G. N. Parkinson, M. P. Lee and S. Neidle, *Nature*, 2002, **417**, 876.
25 E. Herrero-Galan, M. E. Fuentes-Perez, C. Carrasco, J. M. Valpuesta, J. L. Carrascosa, F. Moreno-Herrero and J. R. Arias-Gonzalez, *J. Am. Chem. Soc.*, 2013, **135**, 122.
26 C. Bustamante, Z. Bryant and S. B. Smith, *Nature*, 2003, **421**, 423.
27 Z. Yu, J. D. Schonhoft, S. Dhakal, R. Bajracharya, R. Hegde, S. Basu and H. Mao, *J. Am. Chem. Soc.*, 2009, **131**, 1876.
28 D. Koirala, S. Dhakal, B. Ashbridge, Y. Sannohe, R. Rodriguez, H. Sugiyama, S. Balasubramanian and H. Mao, *Nat. Chem.*, 2011, **3**, 782.
29 S. Dhakal, Y. Cui, D. Koirala, C. Ghimire, S. Kushwaha, Z. Yu, P. M. Yangyuoru and H. Mao, *Nucleic Acids Res.*, 2013, **41**, 3915.
30 M. de Messieres, J. C. Chang, B. Brawn-Cinani and A. La Porta, *Phys. Rev. Lett.*, 2012, **109**, 058101.
31 The change in the end-to-end distance across the transition is force-dependent. The force dependence may be neglected for small differences of force but based on the Worm-like chain model, they will need to be taken into account when extrapolating to zero-force.
32 J. D. Schonhoft, R. Bajracharya, S. Dhakal, Z. Yu, H. Mao and S. Basu, *Nucleic Acids Res.*, 2009, **37**, 3310.
33 M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke and J. M. Fernandez, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 3694.

Supplementary DATA FOR:

Mechanical unfolding of long human telomeric RNA (TERRA)

Miguel Garavís,‡[ab] Rebeca Bocanegra,‡[c] Elías Herrero-Galán,[c] Carlos González,*[b] Alfredo Villasante,*[a] and J. Ricardo Arias-Gonzalez*[d]

SUPPLEMENTARY FIGURES

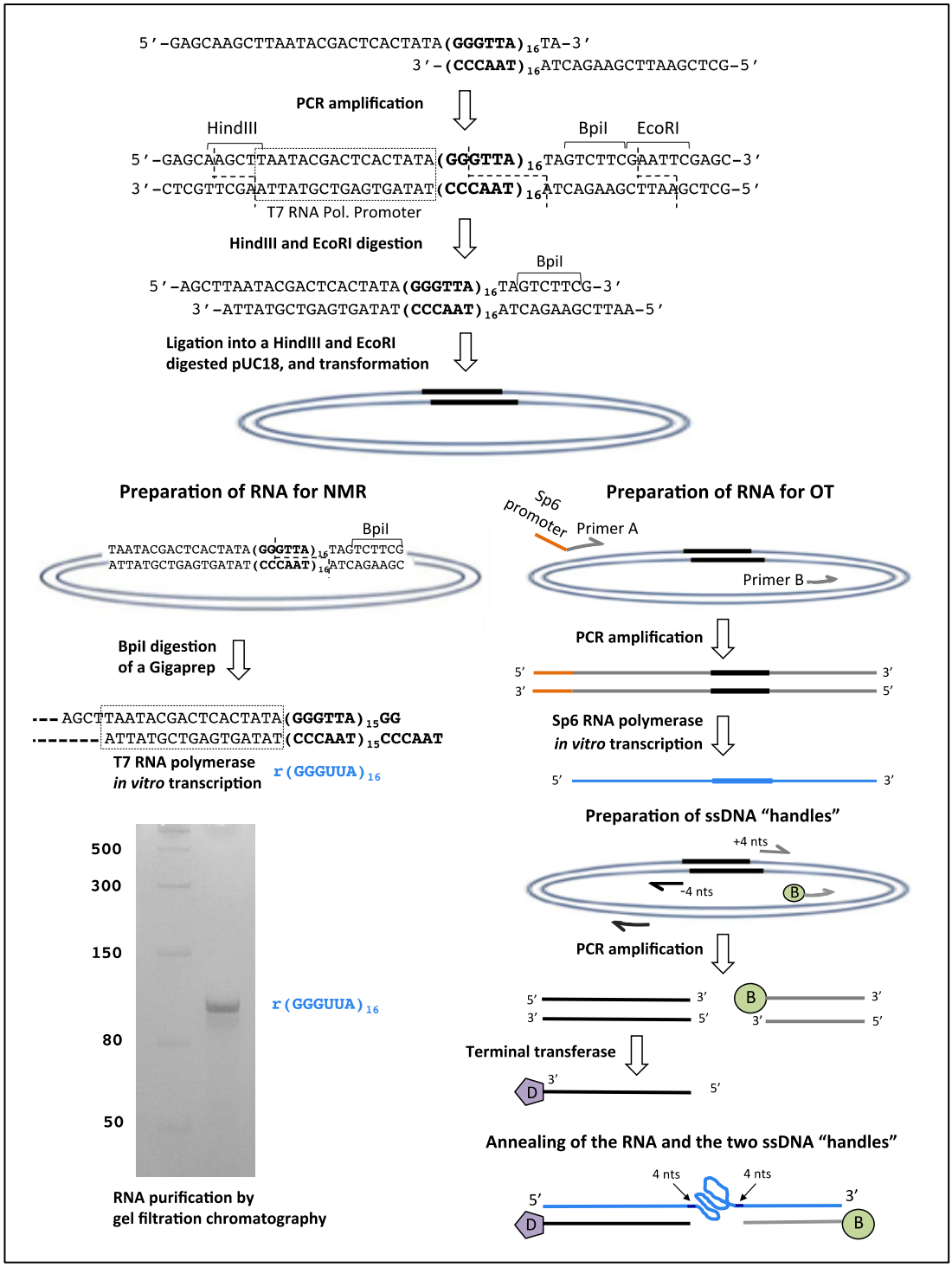


Figure S1. Schematic representation of the steps involved in the preparation of long TERRA samples for NMR experiments and RNA constructions for optical tweezers. RNA appears in blue. For further details see Experimental Methods section.

EXPERIMENTAL METHODS

RNA preparations

The RNA sequences used for NMR, CD and optical-tweezers studies were obtained by *in vitro* transcription. The DNA template was generated by standard PCR methods using the following overlapping primers:
5'-GAGCAAGCTTAATACGACTCACTATA$(GGGTTA)_{16}$TA-3' and
5'-GCTCGAATTCGAAGACTA$(TAACCC)_{16}$TA-3'.[1] The resulting sequence was ligated into the pUC18 vector and transformed into *E. coli* DH5α competent cells. Given the instability of repeated sequences in *E. coli*, different mutations (deletions, amplifications and single point mutations) frequently occur in the telomeric repeats during plasmid cloning. Nevertheless, this disadvantage has been used to obtain plasmid constructions containing variable number of telomeric repeats. These plasmids were then used for the synthesis of the RNA required for NMR of optical-tweezers studies.

To prepare large quantities of RNA for NMR, plasmid DNA was purified from 2 L of cell culture using Qiagen Giga-prep columns. The purified plasmid was resuspended in deionized water at 500 μg/mL and linearized with the Fermentas restriction enzyme BpiI (12 h, 37 ºC, 1 U of enzyme per 50 μg plasmid DNA). The linearized plasmid was used in a 10 mL *in vitro* transcription reaction, where the concentration of $MgCl_2$ and T7 RNA polymerase have been optimized for the highest yield using 25 μL reactions assays. After 2 h at 37 ºC, the reaction was stopped by the addition of EDTA to a final concentration of 2x the concentration of $MgCl_2$, and the T7 polymerase was removed from the mixture by 3 successive extractions with Phenol:Chloroform:Isoamyl Alcohol (25:14:1, v/v) (Sigma Aldrich). The aqueous phase was desalted using 10-DG desalting columns (BioRad) and then purified by FPLC gel filtration chromatography (HILOAD Superdex 75 P 26/60 GE Healthcare Bio-Sciences Corp.). The purified RNA was concentrated and resuspended in 25 mM KPi pH=7 buffer using Amicon Ultra-0.5 mL, 10 kDa centrifugal filters (Millipore).

The optical-tweezers constructions Q1, Q4 and Q6 have the telomeric RNA between two handles made of DNA:RNA hybrid duplexes. The RNA of these constructions was synthesized by *in vitro* transcription using Sp6 RNA polymerase (New England Biolabs) whose promoter sequence was incorporated in the 5' end of the PCR primer. PCR amplification of three pUC18 plasmids containing the desired number of repeats were performed to obtain the DNA templates 1232 bp, 1178 bp and 1112 bp of Q6, Q4 and Q1 transcripts respectively. These DNAs were used, in 25 μL *in vitro* transcription reactions (2 h, 40 ºC) to synthesized RNAs having the telomeric sequences flanked by 677 nt at the 5'end and 405 nt at the 3' end. The DNAs of the handles were obtained by PCR amplification using two pairs of primes. The primers flanking the telomeric repeats start at +4 and -4 nucleotides far from the repeats respectively. The ssDNA of the 3' handle (401 nt) was labelled with biotin using a 5'-biotinylated primer (Integrated DNA Technologies, IDT, Coralville, IA) during PCR amplification. The ssDNA of the 5' handle (673 nt) was labelled at its 3' end with a very short tail of 2-3 digoxigenin-dUTP using DIG Oligonucleotide Tailing Kit, 2nd Generation (Roche) and terminal transferase (Fermentas). The DNA handles where purified by diatomaceous earth. Equimolar amounts of the RNA and the two ssDNA handles were mixed and annealed in a PCR apparatus using the following annealing procedure: 90 ºC 5 min, 85 ºC 10 min, 62 ºC 90 min, 52 ºC 90 min (being the ramp of temperature between steps 0.1 ºC/seg).

Circular Dichroism

CD spectra were recorded on a Jasco 8 J-810 Spectropolarimeter using a 1-mm path-length quartz cuvette. The annealed RNA Q4 was diluted at 2 μM concentration in a volume of 200 μL of 25 mM KPi pH=7 buffer or 25 mM KPi pH=7, 50 mM KCl buffer. Scans from 320 to 220 nm were performed with a 50nm/min scanning speed. For each spectrum, an average of three spectra was taken and the spectrum of the corresponding buffer was subtracted for baseline correction. The melting curves were obtained by recording the change of the molar ellipticity at 263.5 nm in a range of temperatures from 5 ºC to 85 ºC. The temperature was controlled using a Jasco peltier, being the rate of temperature rising 40 ºC/h. The resulting melting temperatures were calculated by fitting the denaturing curves with the program Origin Pro 6.0.

Optical Tweezers

Dual, counterpropagating laser beam (λ = 835 nm) Optical Tweezers were used to measure force from changes in light momentum flux.[2] Single RNA constructs were tethered by opposite hybrid duplex ends between two dielectric polystyrene microspheres: an anti-digoxigenin antibody-coated bead, optically trapped, and a streptavidin-coated bead, held by suction on top of a micropipette. Force on the studied molecule was exerted by moving the micropipette relative to the optically trapped bead through a piezo-controlled stage, and its extension was measured from the distance between the centers of the beads. Stretch−relax cycles were performed at 100 nm/s. All experiments were carried out at room temperature (23 ± 1 °C) in 10 mM Tris Cl (pH 7.8) 100 mM KCl, 1 mM EDTA, in the absence of multivalent ions that may produce condensation.[3-5]

References

(1) Lukavsky, P. J.; Puglisi, J. D. *RNA* **2004**, *10*, 889.
(2) Smith, S. B.; Cui, Y.; Bustamante, C. *Methods Enzymol* **2003**, *361*, 134.
(3) Hormeno, S.; Ibarra, B.; Carrascosa, J. L.; Valpuesta, J. M.; Moreno-Herrero, F.; Arias-Gonzalez, J. R. *Biophys J* **2011**, *100*, 1996.
(4) Hormeno, S.; Ibarra, B.; Valpuesta, J. M.; Carrascosa, J. L.; Arias-Gonzalez, J. R. *Biopolymers* **2012**, *97*, 199.
(5) Hormeno, S.; Moreno-Herrero, F.; Ibarra, B.; Carrascosa, J. L.; Valpuesta, J. M.; Arias-Gonzalez, J. R. *Biophys J* **2011**, *100*, 2006.

Artículo 3.

Discovery of selective ligands for telomeric RNA G-quadruplexes (TERRA) through ^{19}F-NMR based fragment screening

Miguel Garavís, Blanca López-Méndez, Álvaro Somoza, Julen Oyarzabal, Claudio Dalvit, Alfredo Villasante, Ramón Campos-Olivas and Carlos González

Descubrimiento de ligandos selectivos para cuádruplexes de guanina de RNA telomérico (TERRA) mediante cribado de compuestos basado en ^{19}F-RMN

Miguel Garavís, Blanca López-Méndez, Álvaro Somoza, Julen Oyarzabal, Claudio Dalvit, Alfredo Villasante, Ramón Campos-Olivas and Carlos González

Los procesos que permiten la elongación de los telómeros en las células cancerosas permiten que éstas puedan dividirse de forma indefinida. El mecanismo más frecuente de elongación de los telómeros en células cancerosas consiste en la acción de la enzima telomerasa, que se encuentra sobre-expresada en el 85-90% de los cánceres. En el 10-15% de los casos, las células cancerosas utilizan un mecanismo denominado ALT (siglas del inglés *Alternative Lengthening of Telomeres*) que se basa en procesos de recombinación.

Se ha observado que el empleo de moléculas que interaccionan y estabilizan cuádruplex de guaninas teloméricos provoca la disrupción del telómero y en consecuencia la muerte de la célula. Por ello, los cuádruplex de guaninas de TERRA son dianas terapéuticas muy atractivas ya que su presencia se requiere en la formación de la heterocromatina telomérica, tanto en células cancerosas que usan telomerasa para alargar sus telómeros, como en aquellas que usan el mecanismo ALT. Además, TERRA adopta estructuras de orden superior que pueden presentar sitios de unión específicos, aprovechables en el diseño de nuevos fármacos para aumentar su especificidad.

Los métodos que permiten buscar compuestos que interaccionan con cuádruplex de guaninas son escasos y han dado como resultado ligandos muy similares entre sí y con propiedades poco adecuadas para ser utilizados como fármacos. En este contexto, la utilización de estrategias como FBDD (siglas del inglés *Fragment-Based Drug Discovery*) supone una alternativa para el descubrimiento de nuevos compuestos con capacidad para interaccionar con cuádruplex de guaninas. El fundamento de FBDD es la búsqueda de pequeños compuestos que interaccionen con baja afinidad con la diana de interés. Limitar el tamaño de los compuestos que forman parte de este tipo de búsqueda tiene la ventaja principal de reducir el espacio químico en el que se trabaja. Es decir, el número total de compuestos que no superen una masa molecular determinada es mucho menor que si no se aplica dicha restricción. De manera que es posible explorar una mayor proporción del espacio químico usando un número menor de compuestos. Los pequeños compuestos, también llamados fragmentos, que interaccionan con la molécula de interés pueden ser utilizados como puntos de partida para el posterior desarrollo de compuestos de mayor tamaño y afinidad.

La RMN es una técnica capaz de detectar interacciones débiles entre moléculas por lo que su uso es muy adecuado para su aplicación en estrategias de FBDD. La RMN basada en la detección de ^{19}F tiene una serie de ventajas añadidas: (i) permite obtener espectros en los que no aparecen señales del disolvente, del buffer o de otros componentes de la disolución que no contengan flúor en su estructura; (ii) las señales de ^{19}F son finas y aparecen en un intervalo amplio de desplazamiento químico, lo cual reduce las probabilidades de solapamiento entre señales permitiendo analizar varios compuestos distintos en la misma muestra y (iii) las señales de ^{19}F sufren un notable cambio en su anchura de línea cuando el compuesto fluorado interacciona con otra molécula, siendo dicho efecto mucho más acusado en el flúor que en el protón.

En este trabajo, se presenta la aplicación de metodologías basadas en ^{19}F-RMN para la búsqueda de pequeñas moléculas con afinidad por cuádruplex de guaninas de RNA. En concreto, nuestra diana es una molécula de RNA telomérico compuesta por 16 repeticiones de la secuencia del telómero humano. Además, se describen los procedimientos llevados a cabo para la validación de los resultados obtenidos en los experimentos de búsqueda de ligandos (*screening*) y para el análisis de su selectividad por cuádruplex de guaninas. Por último, se presentan el estudio de las interacciones de los compuestos con una secuencia de TERRA más corta, en el cual se explora el modo de interacción, la afinidad de los diferentes ligandos y el efecto que éstos tienen sobre la estabilidad de dicho RNA telomérico.

Aportación personal al trabajo: Obtención del RNA telomérico de 16 repeticiones r(GGGUUA), mediante transcripción *in vitro* y posterior purificación por cromatografía de exclusión molecular. Preparación de las muestras de RNA y DNA para su análisis mediante RMN y CD. Realización y análisis de los experimentos de ^{19}F-RMN, ^{1}H-RMN y dicroísmo circular. Escritura y discusión del manuscrito.

ACS chemical biology

Articles

pubs.acs.org/acschemicalbiology

Discovery of Selective Ligands for Telomeric RNA G-quadruplexes (TERRA) through ^{19}F-NMR Based Fragment Screening

Miguel Garavís,[†,∥] Blanca López-Méndez,[‡] Alvaro Somoza,[§] Julen Oyarzabal,[‡,#] Claudio Dalvit,[‡,⊥] Alfredo Villasante,[∥] Ramón Campos-Olivas,*[,‡] and Carlos González*[,†]

[†]Instituto de Química Física 'Rocasolano', CSIC, Serrano 119, 28006 Madrid, Spain

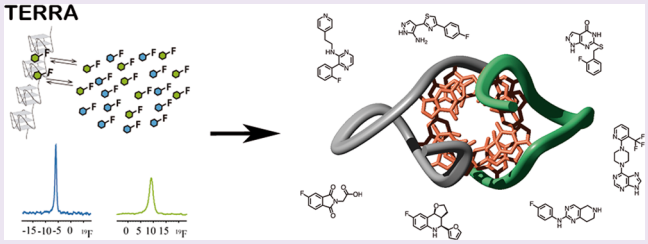
[‡]Spectroscopy and NMR Unit and Experimental Therapeutics Programme, Spanish National Cancer Research Center (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain

[§]IMDEA Nanociencia and CNB-CSIC-IMDEA Nanociencia Associated Unit "Unidad de Nanobiotecnologia", C/Faraday 9, Cantoblanco, 28049 Madrid, Spain

[∥]Centro de Biología Molecular "'Severo Ochoa"' (CSIC-UAM), Universidad Autónoma de Madrid, c/ Nicolás Cabrera1, Cantoblanco, 28049 Madrid, Spain

Supporting Information

ABSTRACT: Telomeric repeat-containing RNA (TERRA) is a novel and very attractive antitumoral target. Here, we report the first successful application of ^{19}F-NMR fragment-based screening to identify chemically diverse compounds that bind to an RNA molecule such as TERRA. We have built a library of 355 fluorinated fragments, and checked their interaction with a long telomeric RNA as a target molecule. The screening resulted in the identification of 20 hits (hit rate of 5.6%). For a number of binders, their interaction with TERRA was confirmed by ^{19}F- and ^{1}H NMR as well as by CD melting experiments. We have also explored the selectivity of the ligands for RNA G-quadruplexes and found that some of the hits do not interact with other nucleic acids such as tRNA and duplex DNA and, most importantly, favor the propeller-like parallel conformation in telomeric DNA G-quadruplexes. This suggests a selective recognition of this particular quadruplex topology and that different ligands may recognize specific sites in propeller-like parallel G-quadruplexes. Such features make some of the resulting binders promising lead compounds for fragment based drug discovery.



Telomeres are specialized nucleoprotein structures found at the end of linear chromosomes. In most eukaryotic chromosomes, the telomeric DNA is composed of tandem arrays of short guanine-rich repeated sequences that are synthesized by the telomere-specific reverse transcriptase telomerase.[1] Importantly, these telomeric sequences terminate in 3′ single-strand G-rich overhangs that are able to fold *in vivo* into DNA G-quadruplexes containing cyclic planar arrangements of four hydrogen-bonded guanines that stack on top of each other.[2,3] Telomeres are transcribed from the subtelomeric regions toward chromosome ends into telomeric repeat-containing RNA (TERRA).[4,5] These noncoding RNAs contain subtelomere-derived sequences and an average of 34 r(UUAGGG) repeats.[6] TERRA molecules fold *in vivo* into RNA G-quadruplexes[7] and act as scaffold for the formation of telomeric heterochromatin.[8,9]

Telomerase is expressed in germ-line cells and in most of cancer cells, but 10−15% of cancers lack telomerase activity and use an alternative lengthening of telomeres (ALT) that involves a recombination-based mechanism.[10] TERRA G-quadruplexes are ideal therapeutic targets because they are required for telomere heterochromatin formation in all cancer cells, even in those that do not require telomerase (ALT-positive tumors),[10] and because their higher-order G-quadruplex structures provide specific binding sites to optimize specificity during drug design.[11]

From a drug discovery perspective, the development of robust approaches for finding compounds that bind G-quadruplex structures is of great interest. Until now, most of the G-quadruplex ligands discovered are the result of rational modifications of compounds previously identified as binders of other structures such as triplexes or duplexes.[12−14] The application of high throughput screening platforms has been mainly limited to virtual docking[15−17] and fluorescence resonance energy transfer (FRET)-based methods.[18−20] It is important to underline that most of the current G-quadruplex ligands are polyaromatic compounds with rather poor drug-like properties. Increasing the chemical diversity of G-quadruplex binders is essential for therapeutic strategies based on targeting this kind of structures. The so-called fragment-based drug discovery strategy (FBDD) is very well-suited to discover new ligands.[21,22] FBDD is aimed to find small fragments that can be elaborated into higher complexity compounds and lead to a drug

Received: February 11, 2014
Accepted: May 16, 2014
Published: May 16, 2014

for the given target. Although this strategy has been used for discovery of novel RNA ligands,[23−25] its application to quadruplex structures has been very limited.[26]

NMR spectroscopy is a very valuable technique for drug discovery, especially for screening and hit validation purposes. NMR is able to detect weak intermolecular interactions and does not require chemically modified samples. Detection of weak interactions is essential for the fragment-based drug design strategy.[27,28] NMR methods based on the observation of changes in ^{19}F ligand spin relaxation provide interesting advantages for the identification of starting compounds in the FBDD strategy.[29] First, there is not background ^{19}F signal from solvent, buffer and solution components. Second, fluorine signals are narrow and present very large chemical shift dispersion, allowing the test of many fluorinated compounds simultaneously. Third, the change in spin relaxation upon binding is much more pronounced in the case of ^{19}F signals due to a significant contribution of chemical shift anisotropy (negligible in proton) and chemical exchange (originating from the difference in ^{19}F chemical shift between free and bound state) to the transverse relaxation of ^{19}F (that determines the line width of the signal). For all these reasons, screening methods based on the observation of ^{19}F ligand signal broadening upon binding are of wide use in the pharmaceutical industry. However, this methodology is mainly used for protein targets. Recent advances in the study of ligand-RNA interactions involving ^{19}F-NMR detection with fluorine-modified nucleic acids[30,31] suggested to us that ^{19}F-NMR based ligand screening approaches might be fruitful in the RNA field.

In this paper, we report the first successful application of fragment-based screening to identify ligands that bind to RNA quadruplexes. To our knowledge, this is the first time that a fragment-based screening based on ^{19}F ligand signal detection is applied to a nucleic acid target.

RESULTS AND DISCUSSION

As mentioned above, TERRA G-quadruplexes are very attractive antitumoral targets because they are required for telomere heterochromatin formation in all cancer cells. From a chemical biology perspective, the development of methods for finding chemically diverse binders is of great interest. This is of particular importance in the case of G-quadruplex structures, since most of their known ligands are relatively similar. Here, we describe a ^{19}F-NMR fragment screening methodology adapted for RNA targets (illustrated in Supporting Information (SI) Figure S1). We apply these methods to identify ligands that bind to a long telomeric RNA composed of 16 r(UUAGGG) repeats ($TERRA_{16}$), an RNA molecule similar in size to endogenous TERRA. We describe the procedures to carry out hit validation experiments and to examine the selectivity of the ligands for RNA G-quadruplexes (SI Figure S2). By studying the interaction with a shorter TERRA construction, we have explored the mode of interaction of some of the hits, and found that different ligands may recognize specific sites in propeller-like parallel G-quadruplexes. In summary, we describe below a methodology able to find chemically diverse compounds with the desired properties of affinity and selectivity for RNA targets, such as TERRA.

Screening of a Fluorinated Fragments Library for Binding to Long Telomeric RNA by ^{19}F-NMR. Long telomeric RNA, $TERRA_{16}$, was prepared at large scale following recently published protocols.[32] $TERRA_{16}$ folding into a structure composed of four consecutive G-quadruplexes was confirmed by circular dichroism (CD) and NMR spectroscopy (SI Figure S3).

A small fragment-oriented library of 355 chemically diverse fluorinated compounds was used for the primary screening. Compounds were selected according to criteria explained in the experimental section. The library contains two types of compounds; those having a trifluoromethyl group, named as CF3, and those having only one fluorine atom, named as CF. Compounds are grouped in 46 cocktail samples containing 8 fluorinated molecules each. The cocktails are composed of molecules from the same group, giving rise to 32 CF3 and 14 CF cocktails, with each compound present at 20 and 50 μM, respectively.

The first step of the screening process is the acquisition of two ^{19}F-NMR spectra of the cocktail samples, a regular 1D, and a 1D containing a CPMG T_2 filter of 200/400 ms (for CF/CF3) that results in a decrease of signal intensity due to relaxation during this time.[29] Most of the fluorinated compounds contain a single CF or CF3 moiety, and consequently each of the signals observed in the spectrum correspond to one compound of the mixture (Figure 1a). Note the lower intensity of the signals in the

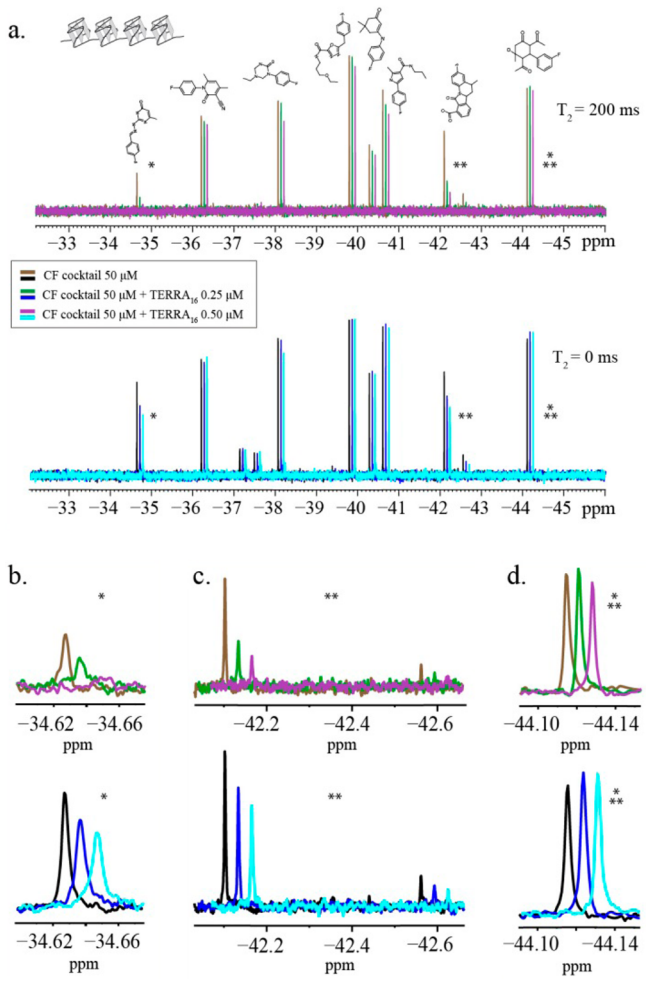


Figure 1. (a) ^{19}F-NMR spectra of a cocktail sample of 8 fluorinated compounds recorded with a T_2 filter of 200 ms (top) and without T_2 filter (bottom). (b) Detail of a signal affected by the addition of $TERRA_{16}$. (c) Detail of two signals corresponding to one compound (purchased as a racemic mixture), both affected by the presence of the RNA target. (d) Enlarged view of a signal not affected by the addition of $TERRA_{16}$. Color codes are described in the inset of panel a. The spectra after the first and second additions have been slightly right-shifted to facilitate visualization. Chemical shifts are referenced to TFA at 0 ppm.

dx.doi.org/10.1021/cb500100z | ACS Chem. Biol. 2014, 9, 1559−1566

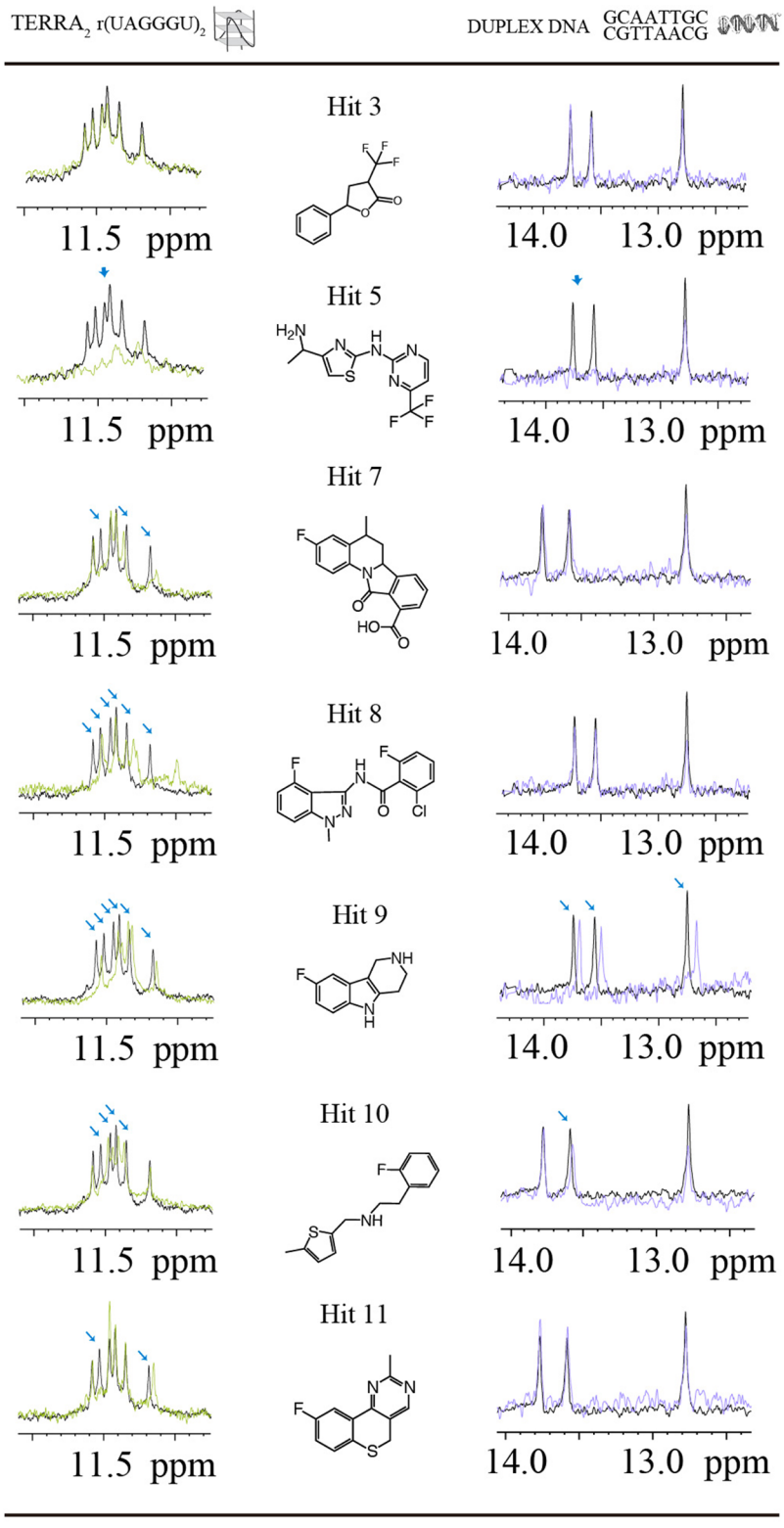


Figure 2. 1D ^{1}H NMR spectra showing the imino regions of TERRA$_2$ and duplex d(GCAATTGC) at 100 μM (except for Hit 8 and Hit 11, for which 33 μM TERRA$_2$ was used) oligo concentration (black). Superimposed NMR spectra of the mixtures containing TERRA$_2$ and hits at near saturating concentrations (see Supporting Information) are shown in the left panels in green and those of mixtures of duplex and 8-fold excess of hit compounds are represented in violet in the right panels. Thin blue arrows mark the imino signals most affected by the interaction.

top spectra, where a T_2 filter was applied. Once the spectra of cocktail samples are acquired, the target macromolecule is added to the mixture in a fragment:TERRA$_{16}$ ratio of 100:1 and the same two NMR spectra are recorded. Upon addition of this first amount of RNA, signals of binders show an increase of their line width (Figure 1b and c), concomitant with a further intensity drop in the T_2 filtered spectrum, while signals from nonbinding compounds remain unmodified (Figure 1d). A second 1:100

dx.doi.org/10.1021/cb500100z | ACS Chem. Biol. 2014, 9, 1559−1566

Table 1. Summary of the Different Nucleic Acid Receptor Binding Results Obtained with the Seven Primary Hits Examined[a]

	RNA				DNA	
	$tRNA^{Phe}$	$TERRA_2$			TEL_2	Duplex
			Δ Tm (°C)	K_D (μM)		
Hit 3	×	×	1.6	—	✓	×
Hit 5	✓	*✓	5.2	—	✓	✓
Hit 7	×	✓	2.5	1259 ± 455	✓	×
Hit 8	×	✓	2.6	121 ± 15	✓	×
Hit 9	×	✓	1.8	1001 ± 68	✓	✓
Hit 10	×	✓	2.8	1932 ± 302	✓	×
Hit 11	×	✓	2.9	—	✓	×
Summary	1/7	6/7			7/7	2/7

[a]Check marks indicate change in the ^{1}H-NMR and/or ^{19}F-NMR spectra in the presence of the ligand, and therefore validation of the hit as a *bona fide* ligand. X indicates no evidence for binding, and − indicates not tested. The asterisk indicates possible aggregation effect. The increase of $TERRA_2$ melting temperature (ΔTm) upon ligand binding (SI Figure S8) as well as the K_D values for the $TERRA_2$:Hit complex formation (SI Figure S7) are also showed. K_D values were determined assuming a single binding site.

equiv of $TERRA_{16}$ is then added, thus reaching a ligand/$TERRA_{16}$ ratio of 50:1. The effect previously described for the NMR signals of the hits should then become more pronounced for compounds affected after the first addition. A few molecules in the cocktails present two signals in the ^{19}F-NMR spectra; in these cases, both signals experience similar effects in the presence of $TERRA_{16}$ (Figure 1c). Following this protocol, screening of the 46 cocktail samples resulted in the identification of 20 molecules whose spectra are affected by the presence of $TERRA_{16}$ (SI Table S1).

Validation of $TERRA_{16}$ primary hits. Validation experiments to confirm the primary hits were performed using samples containing individual compounds. Since some of these experiments require high ligand concentration, and not all hit compounds were available individually, we focused on seven hits. First, the same ^{19}F-NMR experiments were acquired in samples prepared at higher compound concentration than in previous experiments (30/90 μM for CF3/CF compounds), and repeated with three instead of two additions of $TERRA_{16}$, reaching a final ligand/$TERRA_{16}$ ratio of 60:1. All seven compounds tested except one (Hit 3) showed the expected effect. Importantly, line broadening is more pronounced as the amount of $TERRA_{16}$ added to the sample is increased, showing that there is a dose−response effect (SI Figure S4). As in the primary screening experiments, the effect observed in the spectra acquired with relaxation filter is more pronounced. As a second validation experiment, the interaction was confirmed by using independent ^{1}H-NMR techniques. In the case of $TERRA_{16}$, STD experiments (saturation−transfer difference)[33] are suitable for hit validation since there is a significant molecular mass difference between the ligand and the target, and fast dissociation rates are expected (typical for low affinity interactions, $K_D > 1$ μM). All the STD experiments were consistent with the results obtained by ^{19}F-NMR; namely, ligand signals appear in the presence of $TERRA_{16}$ but are missing when $TERRA_{16}$ is absent (SI Figure S5).

Interactions between positive hits and TERRA were also examined by complementary, target-observed NMR methods, by using a shorter TERRA construction. The telomeric RNA sequence 5′-r(UAGGGUUAGGGU)-3′ ($TERRA_2$) was used. The structure of $TERRA_2$ is a dimeric propeller-like parallel G-quadruplex solved by X-ray[34] and NMR spectroscopy.[35] Its 1D NMR spectrum shows six sharp signals between 10 and 12 ppm, which indicates that the sequence is indeed folded into a symmetrical parallel G-quadruplex (Figure 2, left). NMR spectra of the mixtures of each of the seven hits and $TERRA_2$ were initially recorded at two ligand/RNA ratios (16:1 and 8:1). Among the seven ligands analyzed all but one (again Hit 3) induced changes in the NMR spectra of $TERRA_2$ (SI Figure S6). In other case (Hit 5) a general broadening of the RNA signals is observed, most probably due to the formation of RNA aggregates induced by the ligand, and/or chemical exchange. For the remaining five hits changes in the RNA affecting only specific imino signals were detected, suggesting the presence of a well-defined binding site. Evidence for each interaction is also obtained by observing the changes that the complex formation provokes in the fragment ligand signals. Changes in ligand chemical shifts are sometimes combined with a significant line broadening as the amount of $TERRA_2$ increases (SI Figure S6).

The dissociation constants of the $TERRA_2$ interaction with some of the five selective ligands were determined by nonlinear fit of the chemical shift variations of the guanine imino protons upon titration with increasing amounts of each hit compound (see Supporting Information for experimental details and SI

dx.doi.org/10.1021/cb500100z | ACS Chem. Biol. 2014, 9, 1559−1566

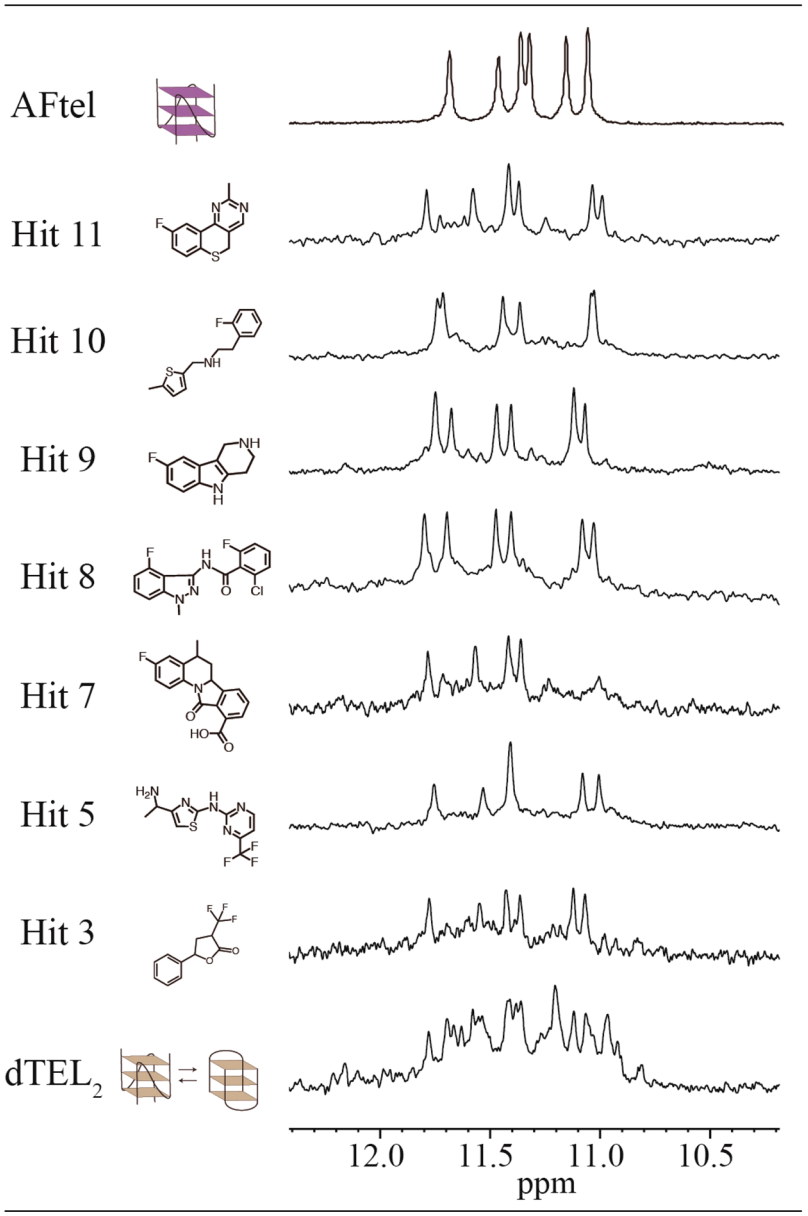


Figure 3. Imino region of the ^{1}H NMR spectrum of isolated dTEL$_2$ (bottom) (at 100 μM), and in the presence of an 8-fold molar excess of each of seven selected hits from the fragment screening of TERRA$_{16}$, whose structures are indicated with the corresponding spectrum. AFtel is a modified analogue of dTEL$_2$, which adopts a single propeller-like parallel quadruplex conformation in solution.[39]

Figure S7). K_D values reported in Table 1 are in the range 120−1900 μM. These affinity constants correspond to ligand efficiency values of 0.17−0.28 kcal/mol/heavy atom, which are appropriate values for initial hits in a FBDD strategy. In addition, CD melting curves of TERRA$_2$ alone and in the presence of the seven selected ligands were obtained (SI Figure S8). All the ligands tested enhanced the thermal stability of TERRA$_2$ by 2 to 5 °C (Table 1). This increase in thermal stability is consistent with the changes observed in the ^{1}H NMR spectra and further confirms the interaction.

Selectivity for G-quadruplexes Structures. Once different ligands of TERRA have been validated, it is crucial to assess their selectivity toward a certain structure before proceeding to the development of more complex ligands with improved properties. To explore their selectivity, ^{19}F- and ^{1}H NMR were used to test ligand binding with other nucleic acids molecules: (i) a tRNA, (ii) a duplex DNA, and (iii) a bimolecular telomeric DNA. First, ligand interactions with other RNA molecules similar to TERRA$_{16}$ in size were explored by ^{19}F-NMR experiments. Phenylalanine tRNA (tRNA^{Phe}) is a suitable target candidate for two reasons: (i) its size (25 kDa), similar to that of TERRA$_{16}$ (31 kDa), and (ii) it possesses a variety of structural motifs that are different from the G-quadruplex,[36] including extensive regions of duplex RNA. We proceeded in a similar way as was done in the hit validation experiments with TERRA$_{16}$. Thus, the variation of the line width of the ligand's fluorine signal is examined upon addition of tRNA^{Phe} to the ligand sample. One of the seven hits assayed was positive for tRNA^{Phe} binding, while the fluorine signals of the other six were not affected by the addition of tRNA^{Phe} (SI Figure S9 and Table 1). This result is a clear indication of the validity and sensitivity of this methodology as a powerful technique to find selective binders.

The B-form DNA helix is the most physiologically common nucleic acids structure. Therefore, the selectivity toward G-

dx.doi.org/10.1021/cb500100z | ACS Chem. Biol. 2014, 9, 1559−1566

quadruplex vs duplex is very frequently evaluated in the characterization of G-quadruplex ligands. The self-complementary sequence 5′-d(GCAATTGC)-3′ was chosen as a suitable duplex model, since its relatively short length and the symmetry of the resulting duplex reduce the spectral complexity. The same receptor-observed ^{1}H NMR experiments performed to study the interaction with $TERRA_2$ were run to examine the possible ligand binding to the duplex. The NMR spectrum of the oligonucleotide at 100 μM concentration in 15 mM KPi buffer shows the expected four imino signals at the characteristic chemical shifts of canonical Watson−Crick base pairs, confirming duplex formation (Figure 2 *right*). Among the seven ligands tested, only two provoked changes in the NMR spectra upon mixing (Table 1). The other five compounds led to very slight or no changes when added to the DNA duplex samples. This result shows that most of the ligands tested recognize and interact specifically with G-quadruplexes and not with DNA duplexes.

Selected Ligands Favor the Parallel Conformation of Telomeric DNA G-quadruplex. Finally, we wondered whether any of the hits were able to interact with a DNA G-quadruplex. Therefore, we decided to use the DNA analogue of $TERRA_2$, 5′-d(TAGGGTTAGGGT)-3′ ($dTEL_2$) to perform the same experiments previously described to validate the binding to $TERRA_2$. $dTEL_2$ is a very convenient quadruplex model that has been extensively studied by either NMR[37] or crystallography.[38] In K^+-containing solution, $dTEL_2$ adopts two interconverting G-quadruplex conformations: a parallel propeller-like fold and an antiparallel fold.[37] These two species are dimeric quadruplexes with different symmetry and can be distinguished easily by counting the number of imino signals in the NMR spectra, since in the propeller-like parallel quadruplex the two subunits are magnetically equivalent, resulting in only six imino signals in the spectrum. The imino region of the NMR spectrum of $dTEL_2$ shows a large number of signals indicating the coexistence of these two conformations (Figure 3, bottom).

Interestingly, this region of the spectra undergoes a dramatic change in the presence of the ligands. The number of imino signals upon complex formation became considerably reduced and corresponds to the parallel propeller-like conformation.[39] Given that these ligands were selected for their ability to interact with RNA quadruplexes, which only adopts a parallel conformation, it seems reasonable that the ligands may favor the formation of the parallel conformation of $dTEL_2$. These results indicate that some of the obtained hits not only are selective for G-quadruplexes vs duplex structures but also exhibit a strong preference for the parallel propeller-like conformation, which is characteristic of telomeric sequences. Moreover, the imino spectra of some of the complexes exhibit significant differences (in particular those of hit5 and hit7), suggesting their binding mode is not the same. This is a very attractive feature in fragment-based drug design projects, where ligand optimization is frequently achieved by combining several initial hits. Preferably, those binding to different sites.

Conclusions. The screening of a fragment-oriented library containing 355 fluorinated compounds has revealed 20 molecules that interact with a long telomeric RNA (hit rate of 5.6%). Their interaction with G-quadruplexes has been further examined on seven hits, confirming that six of them interact with a shorter TERRA construction containing a single G-quadruplex. Interestingly, these hits have a MW < 325, and a chemotype/hit ratio[40] of 1, as each hit bears a different scaffold. Moreover, a subset of the tested compounds with different chemotypes (Hits 7, 8, 10 and 11; hit rate 1.1%) exhibits no interaction with structures distinct to G-quadruplex, such as duplex DNA and tRNA. In addition, these molecules favor the formation of DNA G-quadruplexes with the parallel propeller-like conformation observed in RNA G-quadruplexes, suggesting a selective recognition of this particular quadruplex topology. Therefore, this methodology is able to find chemically diverse compounds with the desired properties of affinity and selectivity for TERRA, suitable as starting points for a fragment-based drug discovery strategy. Finally, it is important to note that the methodology presented here is easily adaptable to other repetitive RNA sequences, as the expanded repeats involved in amyotrophic lateral sclerosis and frontotemporal dementia,[41] and other neurodegenerative diseases.[42]

METHODS

RNA synthesis. The RNA sequence $TERRA_{16}$ $r(GGGUUA)_{16}$ was obtained by *in vitro* transcription[43] and purified following protocols described in previous studies.[32] Short RNA and DNA oligonucleotides, $TERRA_2$ r(UAGGGUUAGGGU) and its DNA analogue $dTel_2$ d(TAGGGTTAGGGT), were synthesized by using phosphoramidite chemistry. Samples for NMR experiments were dissolved in 15 mM KPi pH 7 buffer. Full experimental details are given in the Supporting Information.

Circular Dichroism. CD spectra were recorded on a Jasco 8 J-810 Spectropolarimeter. The annealed $TERRA_2$ and the $TERRA_2$/ligand mixtures were prepared in 200 μL of 15 mM KPi pH 7 buffer at 75 μM RNA strand concentration and 1:8 $TERRA_2$/ligand ratio. The melting curves were obtained by recording the change of the molar ellipticity at 263.5 nm in a range of temperatures from 25 to 85 °C. During the melting experiment, complete CD spectra were recorded every 10 °C. The resulting melting temperatures were calculated by fitting the denaturing curves with the program Origin Pro 6.0.

NMR Experiments. All the ^{1}H-NMR spectra were acquired in Bruker spectrometers. ^{19}F-NMR spectra were acquired at 700 MHz with a dual fluorine-proton SEF probehead, and robotic arm BACS120. ^{1}H NMR spectra were recorded in spectrometers operating at 600 and 800 MHz, equipped with cryoprobes. All spectra were processed with TOPSPIN 2.1 software. Complete experimental details for NMR experiments are given in the SI.

Library of Fluorinated Fragments. The collection of fluorinated molecules for fragment-based screening using ^{19}F-NMR was assembled using three different sources: (1) 113 compounds were selected using the LEF (Local Environment of Fluorine) approach[44] and purchased individually from different vendors (i.e., Maybridge, Sigma-Aldrich, ChemDiv, Enamine, and Asinex), (2) compounds available in house at our institution, and (3) selected compounds from commercially available collections.

The CNIO compounds collection was formed by 41 820 molecules, proprietary and from commercial suppliers, where 6479 compounds were fluorinated. A representative selection of molecules bearing CF_3 and CF was prepared to perform ^{19}F-NMR screening. Considering that the final aim of an initial screening is not only the identification of pharmacological tools for target validation but also discovery of optimal starting points for a plausible medicinal chemistry program, compounds were selected on the basis of two criteria: (i) fragment- or drug-like profile and (ii) coverage of the available chemical space. In addition, taking advantage of the sensitivity of the ^{19}F-NMR approach to identify weak binders, fragment-like molecules (MW < 300)[45] were also favored. Fragment hits are typically molecules with a large ligand efficiency (LE) index[46,47] (calculated as binding energy or IC_{50} vs MW ratio), and are preferred over larger compounds with similar affinity or inhibition activity, because of their potential for rapid expansion and optimization.

The first step in the fluorinated compound selection was filtering out those molecules that do not fit the following criteria (*in-silico* profiling): cLogP < 4, polar surface area (PSA) < 120, number of rotatable bonds < 7, solubility > 100 μM at pH 7.4,[40] and rule of five (Lipinski rules)[48,49] compliance. Ninety nine fragments fulfilling these criteria and bearing CF_3 were selected from the CNIO compounds. On the other hand, from

the drug-like scenario (500 > MW > 300), we had 217 compounds bearing CF_3 and fitting these requirements; however, not all of them were selected. A diversity analysis was performed to select a set of representative compounds covering the available chemical space. This analysis involved clustering those 217 molecules and then selecting a representative compound from each cluster. The clustering approach used is a relocation method based on maximal dissimilarity partitioning, as implemented in Pipeline Pilot.[50] Tanimoto was utilized to measure distance between compounds. Chemical structures were described by circular molecular fingerprints using ECFP_6 descriptors,[51−53] which defines the molecular structure using radial atom neighborhoods. Molecules selected as representatives from each of the 45 clusters covering the analyzed chemical space were also biased by their molecular weight (MW); in fact, less than 10% of selected molecules had a MW > 400 and 69% of those 45 compounds had MW < 350. In a second step, we identified 488 compounds from the CNIO collection bearing only a fluorine atom and fitting the requirements described above. We performed a diversity analysis to select a representative set of 66 compounds covering the chemical space defined by those 488 fragments.

Finally, 32 additional fluorinated compounds from a commercially available collection (from ACD blocks) were purchased to cover additional diversity in the chemical space. The same analysis described above was performed in this set of compounds.

In summary, 355 fluorinated compounds fulfilling high quality standards were selected to build a fragment oriented library for screening by ^{19}F-NMR: 113 selected using the LEF approach and purchased individually from different vendors, 210 from our institutional collection, and 32 from a commercially available collection. 244 CF_3 and 111 CF containing fragments were selected to cover a large chemical space and fluorine local environment. 78.6% of the compounds in the final fluorinated fragment-oriented collection are fragments (MW < 300 Da); a very small number of molecules, 6.5%, with MW > 350.

■ ASSOCIATED CONTENT

Supporting Information

Details of experimental procedures, additional ^{19}F and ^{1}H NMR spectra, CD experiments and melting curves. This material is available free of charge *via* the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Email: cgonzalez@iqfr.csic.es.
*Email: rcampos@cnio.es.

Present Addresses

[⊥]Neuchâtel Platform of Analytical Chemistry, University of Neuchâtel, 2000 Neuchâtel, Switzerland.
[#]Center for Applied Medical Research, CIMA, University of Navarra, Av. Pio XII 55, 31008 Pamplona, Spain.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge Dr. D. V. Laurents for revision of the manuscript and Misses B. de Pablos and D. Velázquez for their excellent technical assistance. We also acknowledge financial support from MICINN (CTQ2010-21567-C02-02, BFU2011-30295-C02-01, SAF2010-15440), *Comunidad Autónoma de Madrid* (S2010-BMD-2457, BIPEDD2), Institutional grant from the Fundación Ramón Areces to the Centro de Biología Molecular "Severo Ochoa"). M.G. was supported by the FPI fellowship BES-2009-027909.

■ REFERENCES

(1) Blackburn, E. H. (1992) Telomerases. *Annu. Rev. Biochem. 61*, 113−129.
(2) Paeschke, K., Simonsson, T., Postberg, J., Rhodes, D., and Lipps, H. J. (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures *in vivo*. *Nat. Struct. Mol. Biol. 12*, 847−854.
(3) Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem. 5*, 182−186.
(4) Azzalin, C. M., Reichenbach, P., Khoriauli, L., Giulotto, E., and Lingner, J. (2007) Telomeric repeat-containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science 318*, 798−801.
(5) Schoeftner, S., and Blasco, M. A. (2007) Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nat. Cell Biol. 10*, 228−236.
(6) Porro, A., Feuerhahn, S., Reichenbach, P., and Lingner, J. (2010) Molecular dissection of telomeric repeat-containing RNA biogenesis unveils the presence of distinct and multiple regulatory pathways. *Mol. Cell. Biol. 30*, 4808−4817.
(7) Xu, Y., Suzuki, Y., Ito, K., and Komiyama, M. (2010) Telomeric repeat-containing RNA structure in living cells. *Proc. Natl. Acad. Sci. U.S.A. 107*, 14579−14584.
(8) Deng, Z., Norseen, J., Wiedmer, A., Riethman, H., and Lieberman, P. M. (2009) TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. *Mol. Cell 35*, 403−413.
(9) de Silanes, I. L., d'Alcontres, M. S., and Blasco, M. A. (2010) TERRA transcripts are bound by a complex array of RNA-binding proteins. *Nat. Commun. 1*, 33.
(10) Bryan, T., Englezou, A., Gupta, J., Bacchetti, S., and Reddel, R. (1995) Telomere elongation in immortal human cells without detectable telomerase activity. *EMBO J. 14*, 4240.
(11) Petraccone, L. (2013) Higher-Order Quadruplex Structures, In *Quadruplex Nucleic Acids* (Chaires, J. B., and Graves, D., Eds.), pp 23−46, Springer, Berlin/Heidelberg.
(12) Guo, Q., Lu, M., Marky, L. A., and Kallenbach, N. R. (1992) Interaction of the dye ethidium bromide with DNA containing guanine repeats. *Biochemistry 31*, 2451−2455.
(13) Haq, I., Ladbury, J. E., Chowdhry, B. Z., and Jenkins, T. C. (1996) Molecular anchoring of duplex and triplex DNA by disubstituted anthracene-9, 10-diones: Calorimetric, UV melting, and competition dialysis studies. *J. Am. Chem. Soc. 118*, 10693−10701.
(14) Collie, G. W., and Parkinson, G. N. (2011) The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chem. Soc. Rev. 40*, 5867−5892.
(15) Cosconati, S., Marinelli, L., Trotta, R., Virno, A., Mayol, L., Novellino, E., Olson, A. J., and Randazzo, A. (2009) Tandem application of virtual screening and NMR experiments in the discovery of brand new DNA quadruplex groove binders. *J. Am. Chem. Soc. 131*, 16336−16337.
(16) Largy, E., Saettel, N., Hamon, F., Dubruille, S., and Teulade-Fichou, M.-P. (2012) Screening of a chemical library by HT-G4-FID for discovery of selective G-quadruplex binders. *Curr. Pharm. Des. 18*, 1992−2001.
(17) Pagano, B., Cosconati, S., Gabelica, V., Petraccone, L., De Tito, S., Marinelli, L., La Pietra, V., Saverio di Leva, F., Lauri, I., and Trotta, R. (2012) State-of-the-art methodologies for the discovery and characterization of DNA G-quadruplex binders. *Curr. Pharm. Des. 18*, 1880−1899.
(18) Renčiuk, D., Zhou, J., Beaurepaire, L., Guédin, A., Bourdoncle, A., and Mergny, J.-L. (2012) A FRET-based screening assay for nucleic acid ligands. *Methods 57*, 122−128.
(19) Benz, A., Singh, V., Mayer, T. U., and Hartig, J. S. (2011) Identification of novel quadruplex ligands from small molecule libraries by FRET-based high-throughput screening. *ChemBioChem. 12*, 1422−1426.
(20) Rahman, K. M., Tizkova, K., Reszka, A. P., Neidle, S., and Thurston, D. E. (2012) Identification of novel telomeric G-quadruplex-targeting chemical scaffolds through screening of three NCI libraries. *Bioorg. Med. Chem. Lett. 22*, 3006−3010.

(21) Hajduk, P. J., and Greer, J. (2007) A decade of fragment-based drug design: Strategic advances and lessons learned. *Nat. Rev. Drug Discovery 6*, 211−219.
(22) Murray, C. W., and Rees, D. C. (2009) The rise of fragment-based drug discovery. *Nat. Chem. 1*, 187−192.
(23) Chen, L., Cressina, E., Leeper, F. J., Smith, A. G., and Abell, C. (2010) A fragment-based approach to identifying ligands for riboswitches. *ACS Chem. Biol. 5*, 355−358.
(24) Cressina, E., Chen, L., Abell, C., Leeper, F. J., and Smith, A. G. (2011) Fragment screening against the thiamine pyrophosphate riboswitch thiM. *Chem. Sci. 2*, 157−165.
(25) Chung, F., Tisné, C., Lecourt, T., Dardel, F., and Micouin, L. (2007) NMR-Guided Fragment-Based Approach for the Design of tRNALys3 Ligands. *Angew. Chem., Int. Ed. 46*, 4489−4491.
(26) Nasiri, H. R., Bell, N. M., McLuckie, K. I., Husby, J., Abell, C., Neidle, S., and Balasubramanian, S. (2014) Targeting a c-MYC G-quadruplex DNA with a fragment library. *Chem. Commun. 50*, 1704−1707.
(27) Pellecchia, M., Bertini, I., Cowburn, D., Dalvit, C., Giralt, E., Jahnke, W., James, T. L., Homans, S. W., Kessler, H., and Luchinat, C. (2008) Perspectives on NMR in drug discovery: A technique comes of age. *Nat. Rev. Drug Discovery 7*, 738−745.
(28) Rees, D. C., Congreve, M., Murray, C. W., and Carr, R. (2004) Fragment-based lead discovery. *Nat. Rev. Drug Discovery 3*, 660−672.
(29) Dalvit, C. (2007) Ligand-and substrate-based 19-F NMR screening: Principles and applications to drug discovery. *Prog. Nucl. Magn. Reson. Spectrosc. 51*, 243−271.
(30) Lombès, T., Moumné, R., Larue, V., Prost, E., Catala, M., Lecourt, T., Dardel, F., Micouin, L., and Tisné, C. (2012) Investigation of RNA−ligand interactions by 19F NMR spectroscopy using fluorinated probes. *Angew. Chem. Int. Ed. 124*, 9668−9672.
(31) Kreutz, C., Kählig, H., Konrat, R., and Micura, R. (2006) A general approach for the identification of site-specific RNA binders by 19F NMR spectroscopy: Proof of concept. *Angew. Chem., Int. Ed. 45*, 3450−3453.
(32) Garavis, M., Bocanegra, R., Herrero-Galan, E., Gonzalez, C., Villasante, A., and Arias-Gonzalez, J. R. (2013) Mechanical unfolding of long human telomeric RNA (TERRA). *Chem. Commun. 49*, 6397−6399.
(33) Mayer, M., and Meyer, B. (1999) Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew. Chem., Int. Ed. 38*, 1784−1788.
(34) Collie, G. W., Haider, S. M., Neidle, S., and Parkinson, G. N. (2010) A crystallographic and modelling study of a human telomeric RNA (TERRA) quadruplex. *Nucleic Acids Res. 38*, 5569−5580.
(35) Martadinata, H., and Phan, A. T. (2009) Structure of propeller-type parallel-stranded RNA G-quadruplexes, formed by human telomeric RNA sequences in K+ solution. *J. Am. Chem. Soc. 131*, 2570−2578.
(36) Quigley, G. J., and Rich, A. (1976) Structural domains of transfer RNA molecules. *Science 194*, 796−806.
(37) Phan, A. T., and Patel, D. J. (2003) Two-repeat human telomeric d (TAGGGTTAGGGT) sequence forms interconverting parallel and antiparallel G-quadruplexes in solution: Distinct topologies, thermodynamic properties, and folding/unfolding kinetics. *J. Am. Chem. Soc. 125*, 15021−15027.
(38) Parkinson, G. N., Lee, M. P., and Neidle, S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature 417*, 876−880.
(39) Martín-Pintado, N., Yahyaee-Anzahaee, M., Deleavey, G. F., Portella, G., Orozco, M., Damha, M. J., and González, C. (2013) Dramatic Effect of Furanose C2′ Substitution on Structure and Stability: Directing the Folding of the Human Telomeric Quadruplex with a Single Fluorine Atom. *J. Am. Chem. Soc. 135*, 5344−5347.
(40) Oyarzabal, J., Pastor, J., and Howe, T. J. (2009) Optimizing the performance of in silico ADMET general models according to local requirements: MARS approach. Solubility estimations as case study. *J. Chem. Inf. Model. 49*, 2837−2850.
(41) Fratta, P., Mizielinska, S., Nicoll, A. J., Zloh, M., Fisher, E. M., Parkinson, G., and Isaacs, A. M. (2012) C9orf72 hexanucleotide repeat associated with amyotrophic lateral sclerosis and frontotemporal dementia forms RNA G-quadruplexes. *Scien. Rep. 2*, 1016.
(42) Gatchel, J. R., and Zoghbi, H. Y. (2005) Diseases of unstable repeat expansion: Mechanisms and common principles. *Nat. Rev. Genet. 6*, 743−755.
(43) McKenna, S. A., Kim, I., Puglisi, E. V., Lindhout, D. A., Aitken, C. E., Marshall, R. A., and Puglisi, J. D. (2007) Purification and characterization of transcribed RNAs using gel filtration chromatography. *Nat. Protoc. 2*, 3270−3277.
(44) Vulpetti, A., Hommel, U., Landrum, G., Lewis, R., and Dalvit, C. (2009) Design and NMR-based screening of LEF, a library of chemical fragments with different local environment of fluorine. *J. Am. Chem. Soc. 131*, 12949−12959.
(45) Makara, G. M. (2007) On sampling of fragment space. *J. Med. Chem. 50*, 3214−3221.
(46) Hopkins, A. L., Groom, C. R., and Alex, A. (2004) Ligand efficiency: A useful metric for lead selection. *Drug Discovery Today 9*, 430−431.
(47) Abad-Zapatero, C., and Metz, J. T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discovery Today 10*, 464−469.
(48) Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev. 23*, 3−25.
(49) Lipinski, C. A. (2004) Lead-and drug-like compounds: The rule-of-five revolution. *Drug Discovery Today: Technol. 1*, 337−341.
(50) *Pipeline Pilot*, version 7.5; (2008) Accelrys, San Diego, CA.
(51) Bender, A., and Glen, R. C. (2005) A discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model. 45*, 1369−1375.
(52) Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci. 44*, 1708−1718.
(53) Glem, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S., and Smith, J. (2006) Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs 9*, 199−204.

Electronic Supplementary Information FOR:

Discovery of selective ligands for telomeric RNA G-quadruplexes (TERRA) through ^{19}F-NMR based fragment screening

Miguel Garavís[1,4], Blanca López-Méndez[2], Alvaro Somoza[3], Julen Oyarzabal[2], Claudio Dalvit[2], Alfredo Villasante[4], Ramón Campos-Olivas[2,*] and Carlos González[1,*]

Experimental Section

Supplementary references

Tables and Figures mentioned in the main text

Table S1. Summary of the 20 binders of $TERRA_{16}$ identified by ^{19}F-NMR screening of fragment cocktails.

Figure S1. Scheme illustrating the basis of the ^{19}F-NMR method applied to detect low affinity ligands.

Figure S2. Scheme summarizing the different steps of the protocol.

Figure S3. Circular dichroism spectrum of $TERRA_{16}$ at 5 ºC (A) and imino region of the 1D ^{1}H-NMR spectra of $TERRA_{16}$ at different temperatures (B).

Figure S4. Examples of 1D ^{19}F-NMR experiments confirming the cocktail screening observations with the isolated compound.

Figure S5. STD experiments of the free hits and of the hits in the presence of $TERRA_{16}$.

Figure S6. 1D ^{1}H-NMR spectra showing the changes produced by the hits on the aromatic signals of $TERRA_2$ and duplex DNA.

Figure S7. Plots showing the chemical shift variation of selected imino protons of $TERRA_2$ upon ligand addition and their best non linear fits for determination of disociation constants.

Figure S8. CD spectra and melting curves of the $TERRA_2$:Ligand complexes.

Figure S9. ^{19}F-NMR experiments testing TERRA ligands for binding to $tRNA^{Phe}$.

EXPERIMENTAL SECTION

Sample preparations

The RNA sequence $TERRA_{16}$ r$(GGGUUA)_{16}$ was obtained by *in vitro* transcription[1]. The DNA template was obtained by standard PCR methods using the following overlapping primers:

5'-GAGCAAGCTTAATACGACTCACTATA$(GGGTTA)_{16}$TA-3' and
5'-GCTCGAATTCGAAGACTA$(TAACCC)_{16}$TA-3'.

The resulting sequence was ligated into the pUC18 vector and transformed into DH5α competent cells. The plasmid DNA was purified (from 2 L of cell culture) using Qiagen Giga-prep columns, and resuspended in deionized water at 500 µg/mL. Purified plasmid was linearized with the Fermentas restriction enzyme BpiI (12 h, 37 ºC, 1 U of enzyme per 50 µg plasmid DNA). The restriction reaction mixture containing the linearized plasmid was used as the template of a 10 mL *in vitro* transcription reaction in which the concentration of $MgCl_2$ and T7 RNA polymerase were previously optimized for the highest yield in 25 µL reactions assays. After 2 h at 37 ºC, the reaction was stopped by the addition of EDTA to a final concentration of 2x the concentration of $MgCl_2$, and the T7 polymerase was removed from the mixture by 3 successive extractions with Phenol:Chloroform:Isoamyl Alcohol (25:14:1, v/v) (Sigma Aldrich). The aqueous phase was desalted using 10-DG desalting columns (BioRad) and then purified by FPLC gel filtration chromatography. Purified RNA was concentrated and resuspended in 15 mM KPi pH=7 buffer using Amicon Ultra-0.5 mL, 10 kDa centrifugal filters (Millipore). The concentration of RNA was measured by using Nanodrop 2000 (Thermo Scientific).

The RNA sequence $TERRA_2$ r(UAGGGUUAGGGU) was prepared using a MerMade4 DNA Synthesizer using phosphoramidites (Link Technologies). After solid-phase synthesis, the solid support was transferred to a screw-cap glass vial and incubated at 55 ºC for 4 h with 1.5 mL of ammonia solution (33 %) and 0.5 mL of ethanol. After the vial was cooled on ice the supernatant was transferred by pipet to microcentrifuge tubes and the solid support and the vial were rinsed with water. The combined solutions were evaporated to dryness using an evaporating centrifuge. The solid was resuspended in 15 mM KPi pH 7 buffer.

The transfer RNA $tRNA^{Phe}$ was purchased from Sigma Aldrich (R4018-1MG) and dissolved at 50 µM in a 15 mM KPi pH 7 buffer.

The DNA oligonucleotides dTel2 d(TAGGGTTAGGGT) and the self-complementary sequence d(GCAATTGC) were purchased from Integrated DNA Technologies, IDT, Coralville, IA. The oligonucleotides were disolved in 15 mM KPi pH 7 buffer and used without further purification.

All the sequences with the exception of $tRNA^{Phe}$ were annealed before being used by heating the samples at 90 ºC for 5 minutes and then cooling down to room temperature overnight.

Circular Dichroism

CD spectra were recorded on a Jasco 8 J-810 Spectropolarimeter using a 1-mm path-length quartz cuvette. The annealed $TERRA_2$ and the $TERRA_2$:Ligand mixtures were prepared in 200 µL of 15 mM KPi pH 7 buffer at 75 µM RNA strand concentration and 1:8 $TERRA_2$:ligand ratio. The melting curves were obtained by recording the change of the molar ellipticity at 263.5 nm in a range of temperatures from 25 ºC to 85 ºC. During the melting experiment, complete CD spectra were recorded every 10 ºC. The scans from 320 to 220 nm were performed with a 50 nm/min scanning speed. For each spectrum, an average of three spectra was taken and the spectrum of the corresponding buffer was subtracted for baseline correction. The temperature was controlled using a Jasco peltier, being the rate of temperature rising 10 ºC/h. The resulting melting temperatures were calculated by fitting the denaturing curves with the program Origin Pro 6.0.

Screening of fluorinated fragments and ^{19}F-NMR Experiments

All the ^{19}F-NMR spectra were acquired in a Bruker spectrometer operating at 700 MHz equiped with a dual fluorine-proton SEF probehead, a robotic arm BACS120, and were processed with TOPSPIN 2.1 software.

Before inclusion of selected fluorinated compounds in the fragment library, quality control NMR experiments were performed with the SPAM filter[(2)] for measuring the solubility, identity, purity and aggregation state of the molecules. Only fragments which had a solubility in Phosphate Buffer Saline (PBS) solution >100 μM and passed the filter were included in the library. In parallel, individual examination of each compound allowed assignment of its ^{19}F NMR signal/s. Compounds were then grouped in sets of eight per coktail, separating CF3 and CF-containg compounds. The choice of the compounds in each cocktail is made so that there is no overlap between different ^{19}F signals of the eight molecules present in the cocktail.

The 32 CF_3 and 14 CF cocktail samples for the ^{19}F-NMR screening were prepared at 20 μM and 50 μM, respectively, by diluting pre-mixed cocktail stocks (10 mM each of the eight compounds in each stock) in d_6-DMSO in a final volume of 550 μL of 15 mM KPi pH 7 buffer, containig 10% D_2O. Two ^{19}F-NMR spectra of each cocktail sample were recorded: a regular 1D, and a 1D containing a CPMG T_2 filter of 200/400 ms (for CF/CF3). The same two spectra were recorded after each of the two additions of $TERRA_{16}$ to ratios $TERRA_{16}$:Cocktail compounds of 1:100 and 1:50 respectively. The annealed $TERRA_{16}$ was added from a 100 μM stock solution in 15 mM KPi pH 7 buffer. The results were analyzed by comparing the intensity of each peak in both the set of spectra without T2 filter and with T2 filter. Those compounds whose signals met a decreasing of intensity proportional to the amount of target in the sample were selected as potential interacting compounds.

^{19}F-NMR Validation experiments

a) Experiments with $TERRA_{16}$

The samples of compounds in individual form were prepared at 30 μM (CF3) and 90 μM (CF) in 550 μL of 15 mM KPi pH 7 10% D_2O buffer, from stock solutions at 80 mM in d_6-DMSO. The same two NMR spectra explained above were recorded before and after each of the 3 succesive additions of $TERRA_{16}$ to the sample ($TERRA_{16}$:Ligand ratios after each addition: 1:175, 1:100 and 1:60).

b) Experiments with transfer RNA

The samples of compounds in individual form were prepared at 67 μM (CF3) and 200 μM (CF) in 550 μL of 15 mM KPi pH 7 10% D_2O buffer, from stock solutions at 80 mM in d_6-DMSO. The same NMR experiments previously described were performed on the samples before and after two additions of $tRNA^{Phe}$. The $tRNA^{Phe}$ was added from a 50 μM stock solution in 15 mM KPi pH 7 buffer, getting to tRNA:compound ratios of 1:150 and 1:75 after the first and the second addition respectively.

^{1}H-NMR Experiments

All the ^{1}H-NMR spectra were recorded in Bruker spectrometers operating at 600 MHz and 800 MHz, equipped with cryoprobes and processed with the TOPSPIN 2.0 software.

a) Titration experiments with $TERRA_2$, $dTEL_2$ and Duplex

NMR samples of interacting compounds were prepared at 800 μM in buffer 15 mM KPi pH 7 10% D_2O. 1D-NMR spectra were acquired at 5 ºC and 25 ºC before the addition of the oligo. $TERRA_2$, $dTEL_2$ and Duplex oligonucleotides were added to the hit samples from annealed 1.5 mM stock solutions in 15 mM KPi pH 7 buffer. Two additions of the oligo were performed reaching oligo:compound ratios of 1:16 and 1:8, and 1D NMR spectra were recorded after each addition.

b) STD experiments

The NMR samples were prepared at 30 μM (CF3 compounds) or 90 μM (CF compounds) and $TERRA_{16}$:Ligand ratio of 1:60 in 15 mM KPi pH 7 and 10% D_2O buffer. In the STD experiments, on-resonance irradiation was set to 5.8 ppm or 6.5 ppm (only for hit3 and hit11) and off-resonance irradiation was set to -187.5 ppm, where no $TERRA_{16}$ resonances are present. The experiments were performed at 5 ºC and 25 ºC. The same STDs were acquired for each hit alone at 100 μM concentration in the same buffer. 256 scans were registered in all the STD experiments.

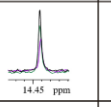
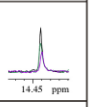
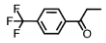
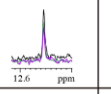
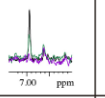
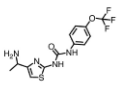
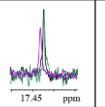
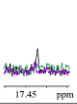
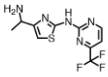
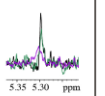
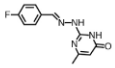
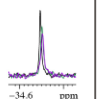
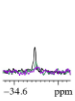
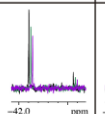
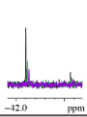
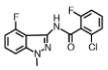
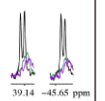
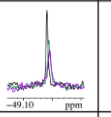
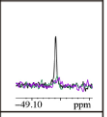
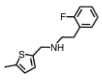
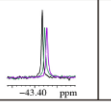
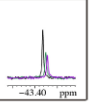
c) NMR determination of the dissociation constants of hit - $TERRA_2$ interactions

The dissociation constants of the interaction between hits 7, 8, 9, 10, and 11 and $TERRA_2$ were determined at 20 ºC following the chemical shift variations of the six guanine imino protons of $TERRA_2$ in the 1D-NMR spectra upon titration with increasing amounts of each hit, up to their approximate solubility limits. Hits 7, 9, and 10 present a relatively high solubility (of at least 0.7, 1.2, and 0.7 mM, respectively) and were titrated over $TERRA_2$ samples at 100 μM concentration, while solubility of hits 8 and 11 is lower (~370 and 80 μM, respectively) and were titrated over samples of $TERRA_2$ at 33 μM. In all cases an initial volume of 240 μL of $TERRA_2$ in 3 mm NMR tubes was used, and effective hit concentrations were recalculated from their proton NMR signals, thus accounting for uncontrollable volume decrease along the titration (10 mM d_6-DMSO stocks of hits were added to $TERRA_2$ solutions previously taken out of the tube). A similar reference titration was also done with addition of the corresponding volumes of d_6-DMSO. Reference NMR spectra of $TERRA_2$ are only slightly affected at 5% DMSO (maximum variation of 0.016 ppm for imino 2, at 11.53 ppm), and only at 15% DMSO larger variations (up to 0.039 ppm for imino 2), indicative of structural rearrangements, were detected. Very reliable and consistent K_D values were obtained by non-linear fit of the chemical shift variations of three different imino signals of $TERRA_2$ upon hit 8 titration (K_D=121 ± 15 μM; Supplementary Figure S7). For the most soluble hit 9 the titration curve could also be properly sampled, thus obtaining a relatively precise K_D=1001 ± 68 μM (Figure S7). In contrast, only very approximate values could be obtained from the changes of the imino 2 signals upon titration of hits 7 and 10 (Figure S7; K_D=1.26 ± 0.46 and 1.93 ± 0.30 mM, respectively). Finally, only a lower limit of 0.3 mM could be estimated for the K_D of the least soluble hit 11, that only resulted in a linear change in the imino 2 signal over the concentration tested (not shown).

SUPPLEMENTARY REFERENCE

1. Makara, G. M. (2007) On sampling of fragment space, *J. Med. Chem. 50*, 3214-3221.
2. Dalvit, C., Caronni, D., Mongelli, N., Veronesi, M., and Vulpetti, A. (2006) NMR-based quality control approach for the identification of false positives and false negatives in high throughput screening, *Curr.Drug Discov.Technol. 3*, 115-124.
3. Vulpetti, A., and Dalvit, C. (2012) Fluorine local environment: from screening to drug design, *Drug Discov Today 17*, 890-897.

SUPPLEMENTARY TABLES

Hit nº	Structure	MW	Effect observed	
			T_2 = 0 ms	T_2 = 200 ms
1		162.11		
2		202.17		
3		230.18		
4		408.40		
5		303.31		
6		246.24		
7		311.31		
8		321.71		
9		190.22		
10		249.35		

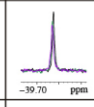
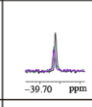
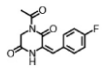
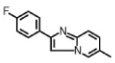
Hit nº	Structure	MW	Effect observed	
			T_2 = 0 ms	T_2 = 200 ms
11		232.28		
12		262.24		
13		226.24		
14		349.31		
15		244.27		
16		276.29		
17		259.28		
18		233.16		
19		260.29		
20		328.77		

Table S1. Summary of the 20 binders of $TERRA_{16}$ identified by ^{19}F-NMR screening of the fragment cocktails. Their molecular structure, molecular mass, and a detailed view of the effect produced by $TERRA_{16}$ addition in the ^{19}F signal are shown. Colour codes for ^{19}F-NMR spectra are: Black, isolated compound at 50 μM (CFs) or 20 μM (CF_3) concentration; Green, first addition of $TERRA_{16}$ (1:100 $TERRA_{16}$:Ligand ratio); and purple, second addition of $TERRA_{16}$ (1:50 $TERRA_{16}$:Ligand ratio). It is worth noting the larger number of CF containing molecule hits when compared to the CF_3 containing molecules (14 *vs* 6) despite the significantly fewer CF containing molecules present in the library (31% CF vs 69% CF3) and the higher concentration (50 vs. 20 μM) used for the screening. This is due to the higher relative sensitivity of the CF containing compounds to receptor binding owing to the larger observed difference in chemical shift of the fluorine signal between the free and bound state and probably also to a larger ^{19}F CSA.[3]

SUPPLEMENTARY FIGURES

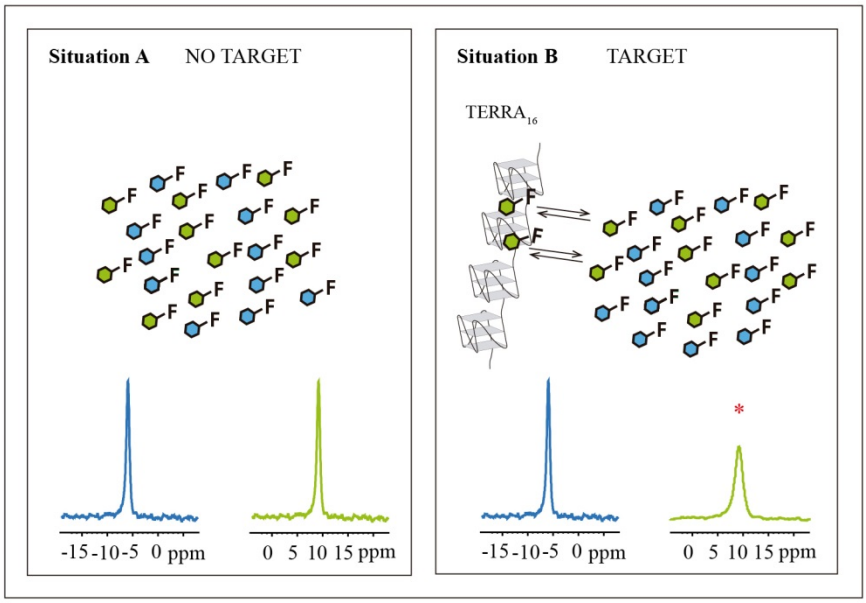


Figure S1 Schematic view of the ^{19}F-NMR method used to detect interactions between small fluorinated ligands and receptor targets. The ligand ^{19}F signal linewidths increase significantly for those compounds that interact with the target. Even a low population of bounded compound provokes observable signal broadening in the ^{19}F NMR spectra.

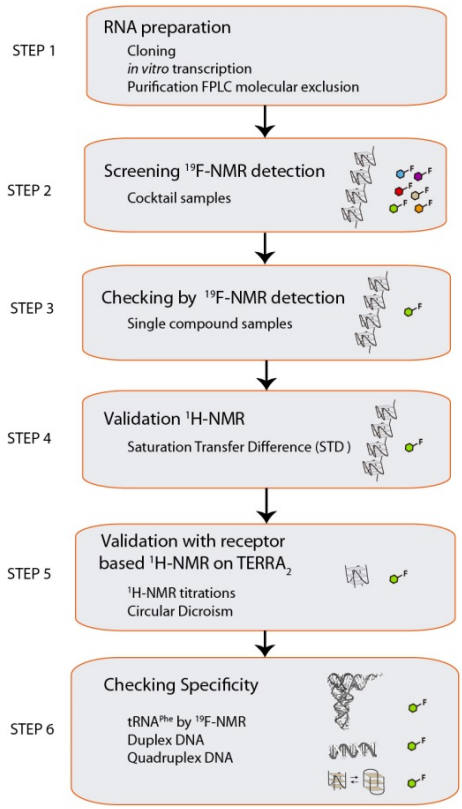


Figure S2. Diagram showing the successive steps carried out for screening and subsequent study of the compounds identified as TERRA ligands.

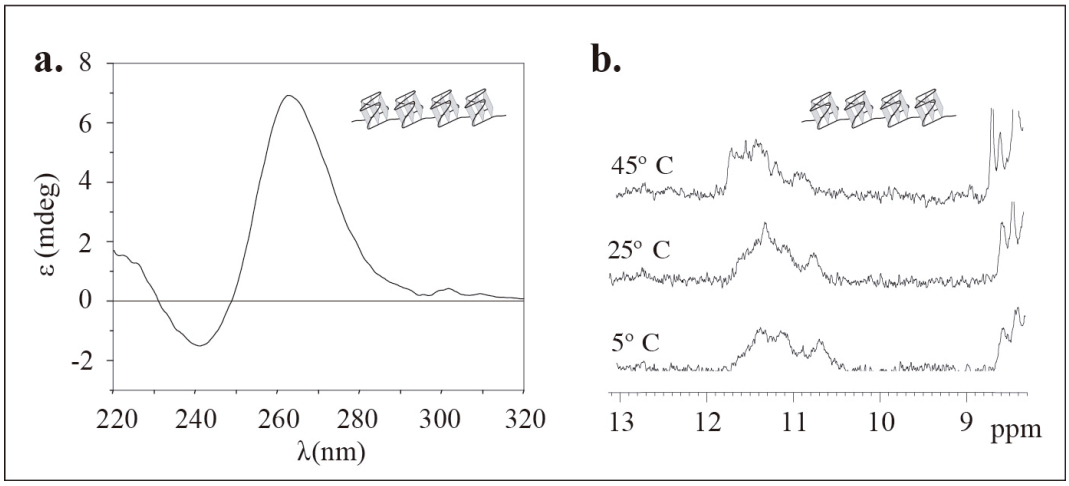


Figure S3. a. CD spectrum of $TERRA_{16}$ at 5 ºC in 15 mM KPi pH 7 buffer and 2 µM oligonucleotide concentration. The presence of a maximum at 260 nm and a minimum at 240 nm is consistent with a parallel conformation. **b.** 1D-^{1}H NMR experiments of 10 µM $TERRA_{16}$ at different temperatures, in 15 mM KPi pH 7 buffer and 10% D_2O. The spectra show imino signals at the characteristic chemical shifts of G-quadruplexes.

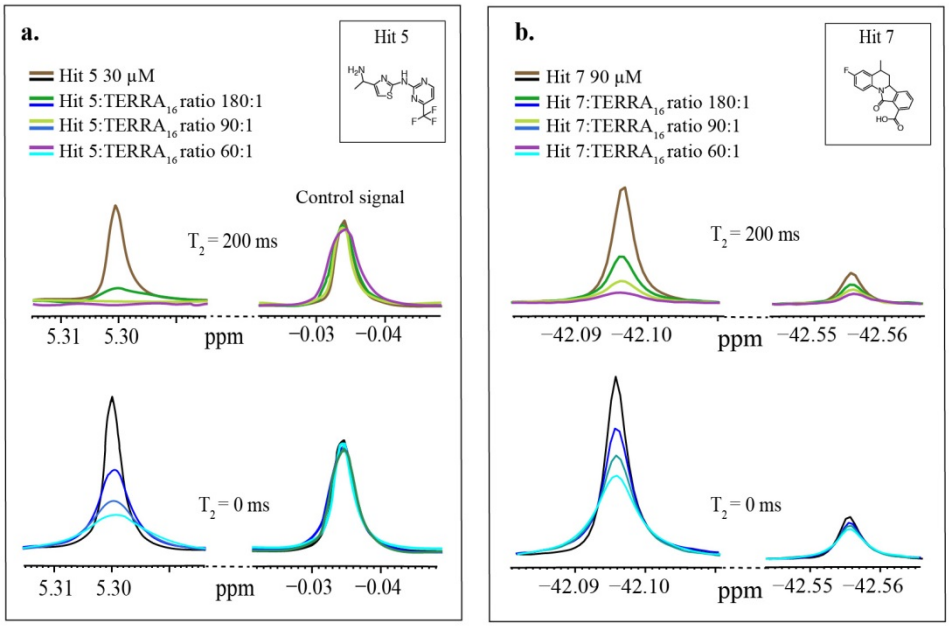


Figure S4. ^{19}F-NMR experiments confirming the interaction of two selected screening hits with $TERRA_{16}$. Three volumes of $TERRA_{16}$ were added to each individual hit sample and the corresponding spectra are shown superimposed (see legend) for **a.** Hit 5 and **b.** Hit 7. The dose-response effects are clearly observed on the fluorine signal corresponding to Hit 5 and on the two signals of compound 7 (in this particular case, the two signals correspond to the two stereoisomers of the Hit 7 present in the racemic mixture of the purchased compound) but no changes are produced on a control signal present in the same spectrum.

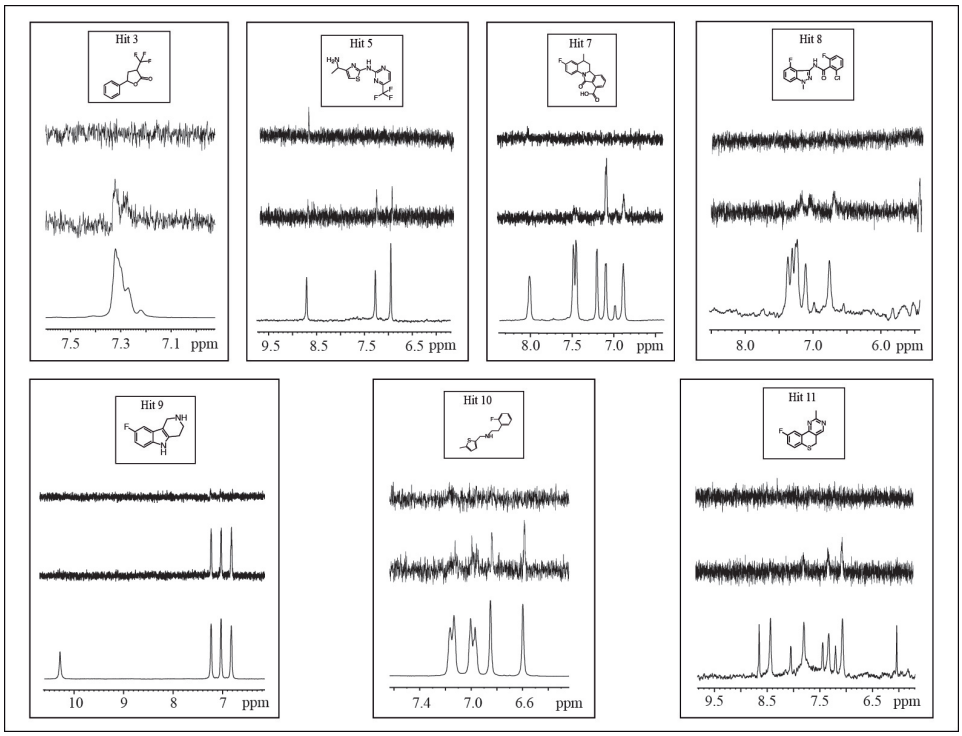
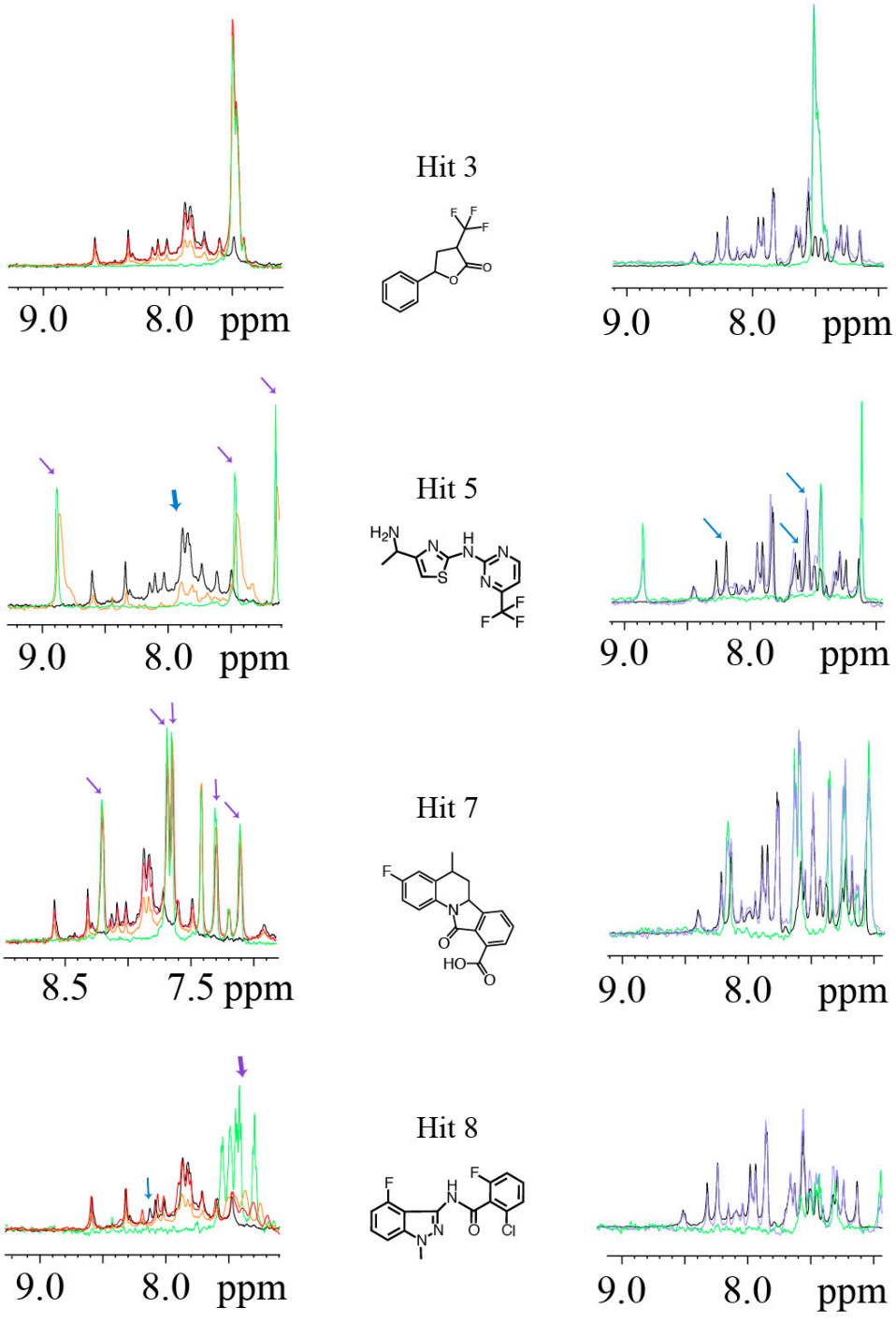


Figure S5. STD experiments confirming the interaction of the seven selected screening hits with $TERRA_{16}$. In each panel the STD spectrum of the indicated compound recorded in the presence (middle spectra) and absence (top spectra) of $TERRA_{16}$ is shown, together with the ^1H 1D NMR spectrum (bottom spectra) of the samples containing a mixture of the compound and $TERRA_{16}$ at 90 μM and 1.5 μM, respectively. Note that signals of $TERRA_{16}$ are not visible at this low concentration. The STD experiment of compound 5 (top) shows the signal of the exchangeable proton at 8.7 ppm probably due to magnetization transfer through water (possibly affected by saturation pulses applyed here at 5.8 ppm). As an example, compound 7 exhibits very large differential STD effects, providing epitope mapping information.


Hit 3
9.0 8.0 ppm
9.0 8.0 ppm
Hit 5
9.0 8.0 ppm
9.0 8.0 ppm
Hit 7
8.5 7.5 ppm
9.0 8.0 ppm
Hit 8
9.0 8.0 ppm
9.0 8.0 ppm

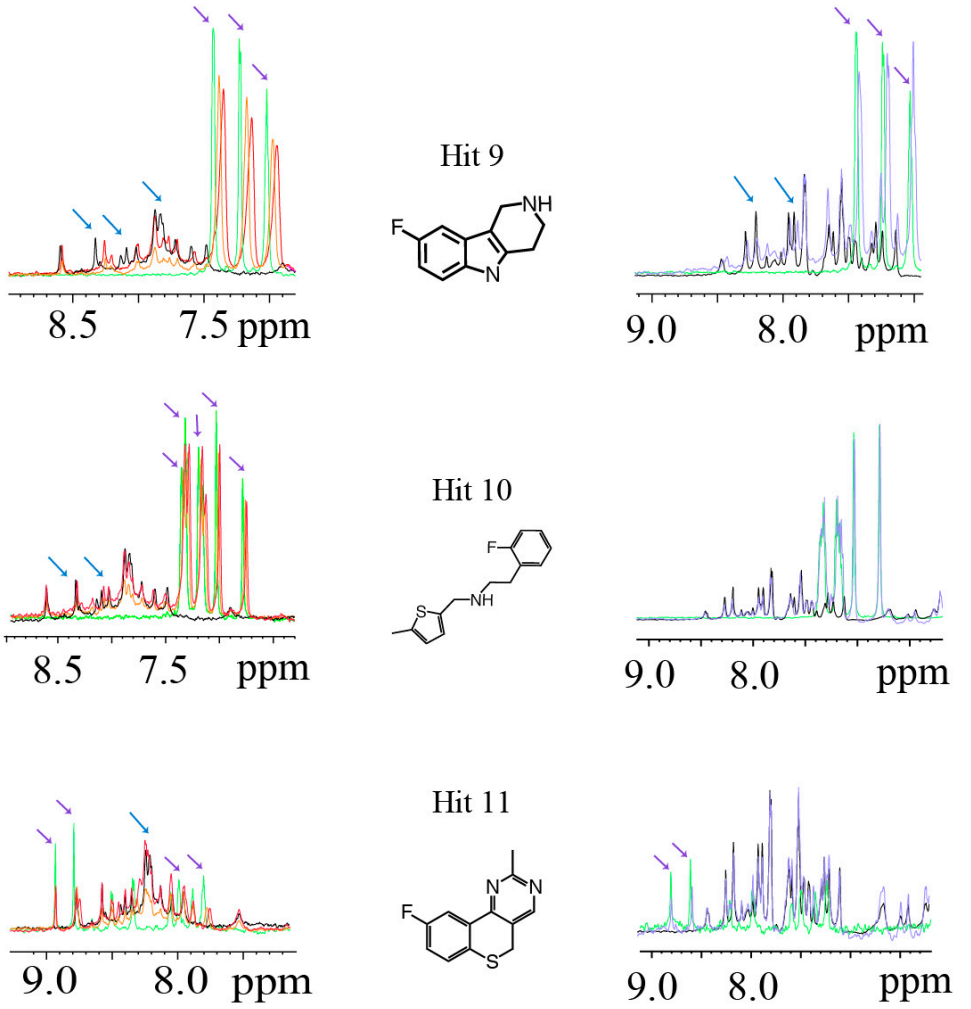


Figure S6. 1D ^{1}H-NMR spectra showing the aromatic region of $TERRA_2$ and duplex d(GCAATTGC) at 100 μM oligo concentration (black), T=25 ºC. Spectra of the free compounds (green) at ~800 μM are also displayed. Superimposed NMR spectra of the mixtures containing $TERRA_2$ at 50 μM (16:1 ligand:$TERRA_2$ ratio) and at 100 μM (8:1 ligand:$TERRA_2$) are shown in the left panels in orange and red, respectively. Mixtures of ligands and duplex at 100 μM are represented in violet in the right panels. Thin blue and purple arrows mark the changes produced by the interaction in the target and ligand signals, respectively. The two wide arrows in hit5 and hit8 panels indicate signal disappearance or very pronounced changes in the spectra as a consequence of the interaction

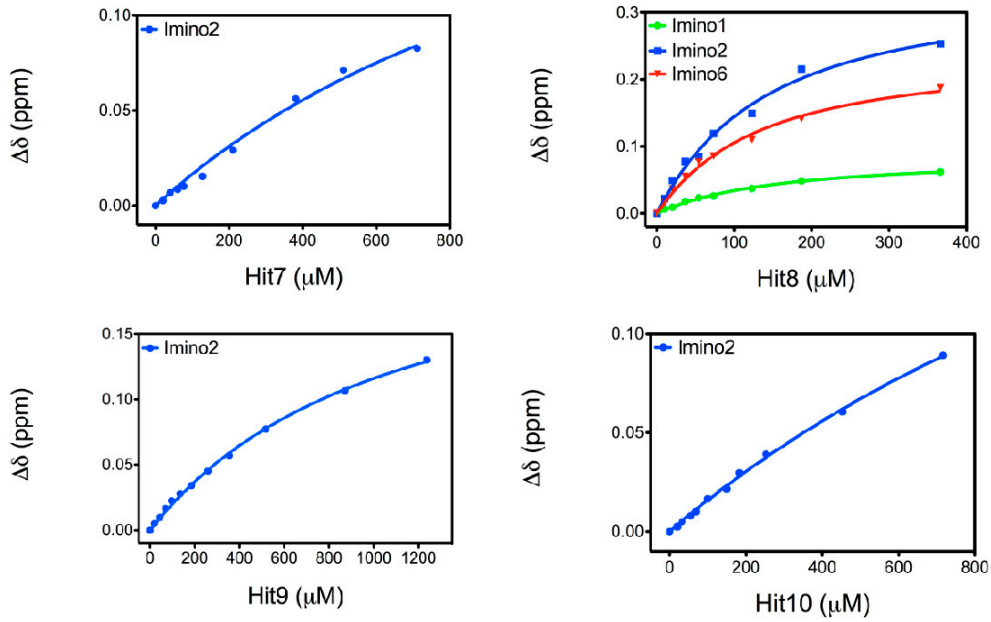


Figure S7. Representation of the chemical shift variation of selected imino protons of $TERRA_2$ upon titration with selected hits, and their best non linear fits for determination of disociation constants. Imino1 to imino6 reffer to the six imino signals detected in the 1D-1H-NMR spectrum of $TERRA_2$ from left to right (Figure 2). Independent fit of three different imino signals of $TERRA_2$ upon hit 8 titration yield K_D values of 130 ± 13 μM (imino1), 121 ± 20 μM (imino2), and 111 ± 15 μM (imino6). See Table 1 for all K_D values.

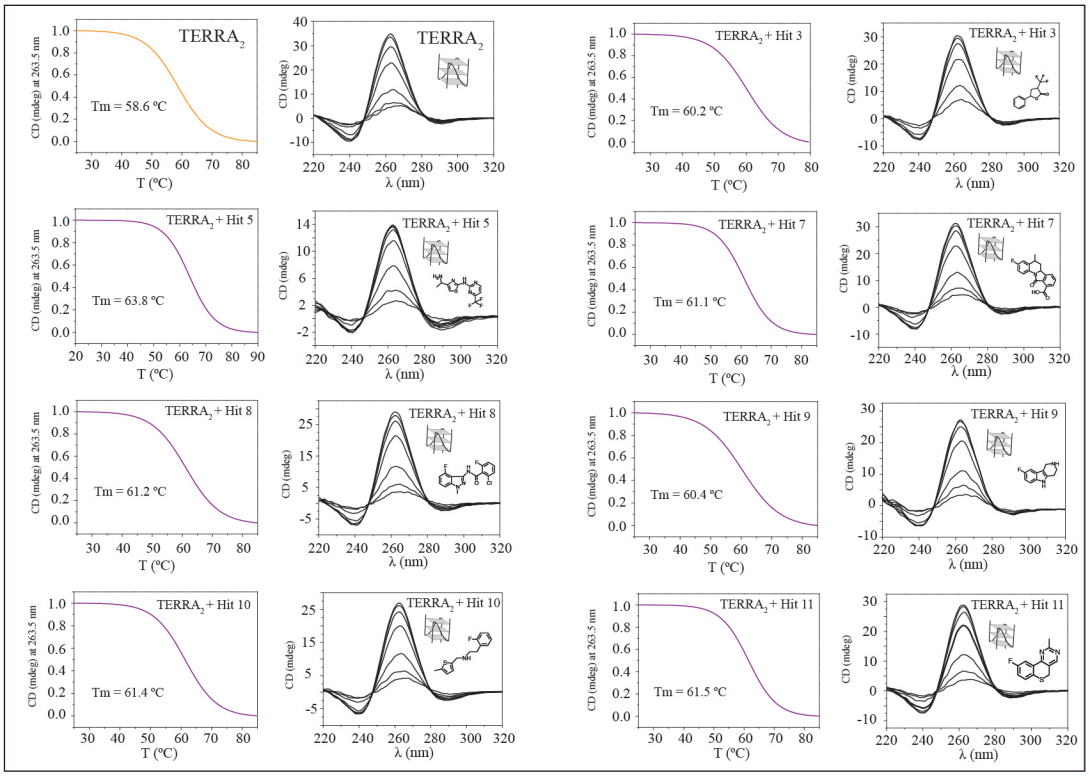


Figure S8. CD melting curves of $TERRA_2$ at 75 μM oligo concentration in 15 mM KPi pH 7 buffer (orange curve) and $TERRA_2$ in conbination with each ligand (violet curves) at a ligand:TERRA2 ratio around 1:8. The melting temperatures are indicated under the melting curves.

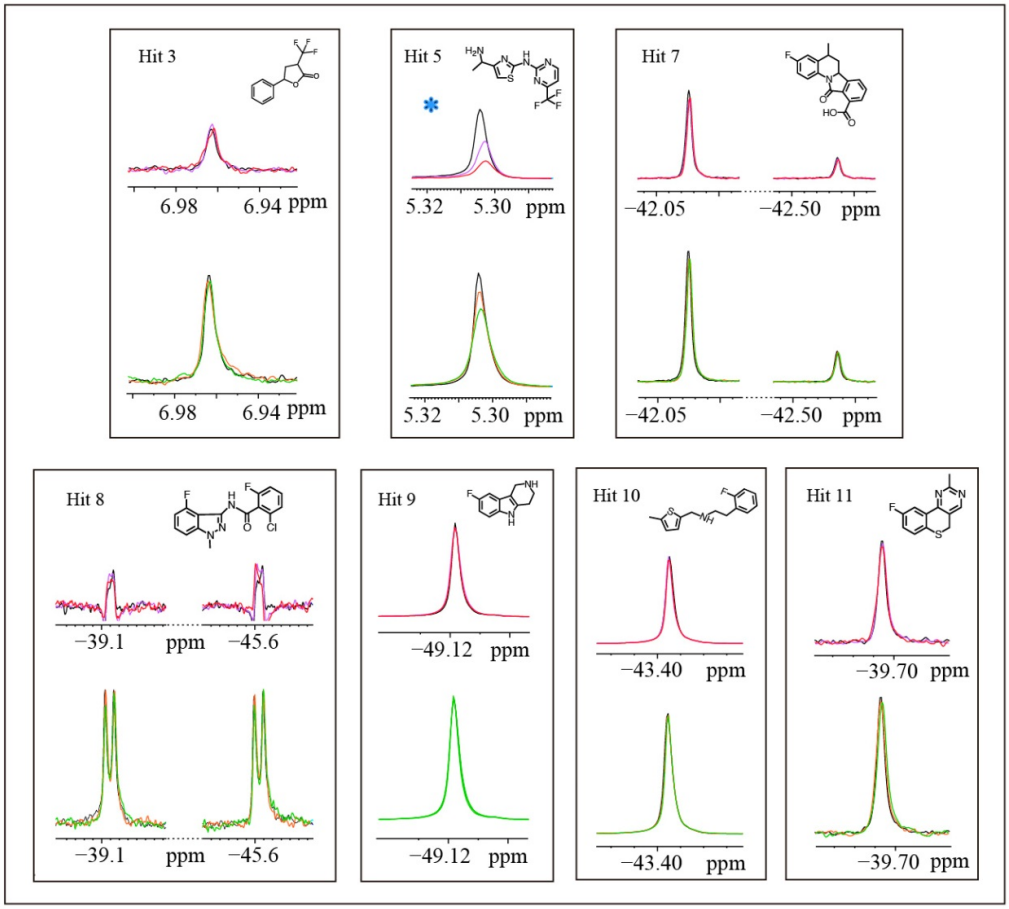


Figure S9. ^{19}F-NMR experiments showing ligand signals before (black) and after the addition of tRNA^{Phe} at ligand:RNA ratios of 150:1 (orange and violet) and 75:1 (green and red). Top spectra in each panel were recorded with a T_2 filter of 200 ms. All the samples were prepared in 15 mM KPi pH 7 buffer and at 200 μM (CF compounds) or 67 μM (for the CF3-containing Hits 3 and 5) ligand concentration. Blue stars indicate the Hit 5 compound, whose signals is affected by the succesive additions of tRNA^{Phe}.

Artículo 4.

Centromeric alpha-satellite DNA adopts dimeric i-motif structures capped by AT Hoogsteen base-pairs.

Miguel Garavís, Núria Escaja, Valerie Gabelica, Alfredo Villasante and Carlos González

El DNA satélite alfoide adopta estructuras *i-motif* diméricas tapadas por pares de bases A:T tipo Hoogsteen

Miguel Garavís, Núria Escaja, Valerie Gabelica, Alfredo Villasante and Carlos González

El centrómero es la región del cromosoma sobre la cual se produce el ensamblaje del cinetocoro, una estructura multiproteíca cuya función es la correcta segregación cromosómica durante la división celular.

El DNA centromérico está constituido por secuencias altamente repetidas denominadas DNA satélite. El satélite alfoide es la secuencia de DNA que se repite en el centrómero humano. Se trata de una secuencia rica en adeninas y timinas de una extensión de 171 pares de bases (pb) que se repite a lo largo del centrómero. Se pueden distinguir dos tipos de secuencias alfoide, el tipo A y el tipo B. La diferencia entre ambas radica en la composición de una región de 17 pb rica en guaninas y citosinas que aparece en cada monómero. Esta región recibe el nombre de A box, en los monómeros tipo A y de B box o CENP-B box en los monómeros tipo B. La denominación CENP-B box deriva del hecho de que esta secuencia es el lugar de reconocimiento de la proteína centromérica CENP-B. Los monómeros tipo B están presentes en los centrómeros de todos los cromosomas excepto en el cromosoma Y, que solo tiene monómeros tipo A.

La estructura de la secuencia CENP-B box fue estudiada por Gallego et al en 1997. Los resultados de RMN mostraron que la hebra rica en citosinas de la secuencia CENP-B box era capaz de formar un *i-motif* a pH ácido. La estructura tridimensional de una versión truncada de la CENP-B box es la de un *i-motif* dimérico con una asociación cabeza-cabeza de ambas hebras y estabilizado por una tétrada de surco menor G:T:G:T. En este trabajo, se ha determinado la estructura en disolución de las dos variantes principales de la secuencia A-box. Los resultados de RMN, CD y espectrometría de masas muestran que las hebras ricas en citosina de dos secuencias A-box se asocian con una orientación cabeza-cola para formar *i-motifs* estabilizados por pares no canónicos $C:C^+$, T:T y A:T Hoogsteen. Las dos variantes de la secuencia A-box forman estructuras similares, con la salvedad de que los residuos que intervienen en la formación de los pares $C:C^+$ son distintos en cada caso. Estos resultados, junto con los obtenidos para el estudio de la secuencia B-box, sugieren que la estructura *i-motif* podría tener un papel en la organización de la cromatina centromérica.

Aportación personal al trabajo: Preparación de las muestras de DNA en diferentes condiciones de pH y concentración, para su posterior estudio por RMN, espectrometría de masas y dicroísmo circular. Adquisición de los experimentos de RMN, espectrometría de masas y dicroísmo circular. Asignación de los espectros de RMN y participación en el cálculo de la estructura. Escritura y discusión del manuscrito.

Centromeric alpha-satellite DNA adopts dimeric i-motif structures capped by AT Hoogsteen base pairs.

Miguel Garavís[1,2], Núria Escaja[3], Valérie Gabelica[4,5], Alfredo Villasante[2*], and Carlos González[1*]

[1]Instituto de Química Física Rocasolano, CSIC, Serrano 119, 28006 Madrid, Spain. [2]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Nicolás Cabrera 1, 28049 Madrid, Spain. [3]Departament de Química Orgànica and IBUB, Universitat de Barcelona, Martí i Franquès 1-11, 08028 Barcelona. [4]Univ. Bordeaux, ARNA Laboratory, IECB, 2 rue Robert Escarpit F-33600 Pessac, France. [5]Inserm, ARNA Laboratory, 146 Rue Leo Saignat, F-33000 Bordeaux, France.

Abstract

Human centromeric alpha-satellite DNA is composed of tandem arrays of two types of 171 bp monomers; type A and type B. The differences between these types are concentrated in a 17 bp region of the monomer called the A/B box. Here, we have determined the solution structure of the two main variants of the human alpha-satellite A box. We show that, under acidic conditions, the C-rich strands of two A boxes self-recognize and form a head-to-tail dimeric i-motif stabilized by four intercalated hemi-protonated C:C$^+$ base pairs. Interestingly, the stack of C:C$^+$ base pairs is capped by T:T and Hoogsteen A:T base pairs. The two main variants of the A box adopt a similar three-dimensional structure, although the residues involved in the formation of the i-motif core are different in each case. Together with previous studies that showed that the B box (known as the CENP-B box) also forms dimeric i-motif structures, our finding of this non-canonical structure in the A box raises the possibility that i-motifs may play a role in the structural organization of the centromere.

Introduction

The centromere is the chromosomal region on which the kinetochore, a large multiprotein structure, forms and attaches to spindle microtubules and ensures proper chromosome segregation. It is now recognized that both genomic competency and epigenetic mechanisms act together to define the site of kinetochore assembly [1-4]. Nucleosomes containing the centromere-specific histone H3 variant CENP-A provide the epigenetic mark to establish the centromere-specific chromatin structure. In particular, the characteristic DNA unwrapping at CENP-A nucleosomes may enable these centromeric particles to adopt the three-dimensional chromatin structure required for centromere function[5,6].

The centromeres of eukaryotic chromosomes are a paradox in that their function is conserved but their tandemly repeated "satellite" DNA sequences evolve rapidly by inevitable recombination processes. However, the rapid and adaptive evolution of CENP-As involves regions that are predicted to contact centromeric DNA[7-9], and this provides compelling evidence that CENP-As evolve in concert with satellite DNAs[10].

Centromeric satellites may have evolved to stabilize centromeric nucleosomes against the

pulling forces they undergo during chromosome segregation. These satellite DNAs may have been selected not by their primary sequence but for their ability to form non-canonical DNA structures[11]. Since this effect may be easily unnoticed by sequence analysis tools, structural studies on satellite DNA sequences are necessary to corroborate this hypothesis.

Since human centromeres have been extensively investigated, their alpha-satellite DNA is a useful model to understand the role of satellite DNAs in centromere structure and function. This satellite DNA has evolved highly homogenized higher-order repeats of the basic 171 bp repeat unit at functional centromeres and divergent monomeric repeats at the periphery (centromeric relics) (Figure 1). In humans, there are two types of alpha-satellite monomers: type A and type B (Figure 1), but lower primates have only type A satellites. The type B monomer evolved from type A in the common ancestor of great apes[12]. The differences between A and B types lie in a 17 bp segment called the A/B box. This clustering is considered to be indicative of positive selection[13]. The B box is also known as the CENP-B box because it is the binding site of the centromeric protein CENP-B. Type B monomers have spread to all human centromeres except the Y chromosome, which only has type A monomers[14]. Similarly, the same 17 bp CENP-B box sequence appears in the mouse centromeric minor satellite DNA and not in the mouse Y centromeric repeat DNA[15,16]. Therefore, this motif has evolved independently in an ancestor of the *Mus musculus domesticus*.

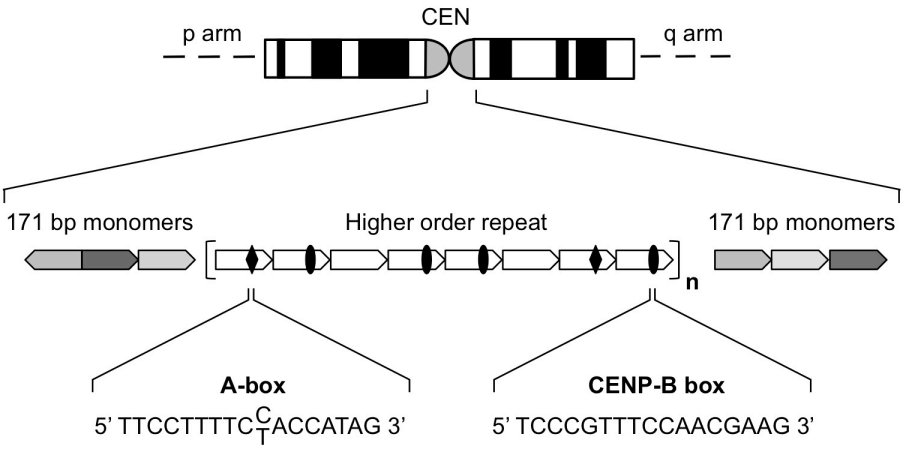


Figure 1. *Structure of the centromeric alpha-satellite DNA showing the position of the A and B boxes.*

It has been proposed that CENP-B may function by juxtaposing two distal CENP-B boxes through DNA-protein and protein-protein interactions[17], but it has also been suggested that the binding of CENP-B may facilitate the folding of CENP-B box C-rich strands into dimeric i-motif structures[18,19]. Subsequent studies have shown that the C-rich strand of the human satellite III DNA can also adopt the i-motif structure[20], but it is now known that this human satellite does not belong to the functional centromere.

The i-motif is a four-stranded intercalated structure formed by the association of two parallel-stranded duplexes connected by hemi-protonated $C:C^+$ base pairs[21,22]. The two duplexes are

intercalated in opposite orientations. Since i-motif formation requires protonation of cytosines[23], these structures are more stable at acidic pH, although, depending on particular C-rich sequences, they can fold close to neutral pH[24]. I-motifs can also exist at neutral pH under molecular crowded conditions[25] and under transcriptionally induced negative superhelicity[26]. The recent remarkable finding that the transcription of centromeric alpha-satellite DNA during mitosis is required for centromere function[27] supports that negative superhelicity conditions may be common in these regions and, consequently, favor i-motif formation under physiological conditions. In the last few years, i-motifs are attracting a great deal of interest for their implication in biological processes and in DNA-nanotechnology (for recent reviews see[28-30]).

In the present study, we determine the solution structure of the A box of the human centromeric alpha-satellite DNA. The structure is a dimeric i-motif resulting from a head-to-tail association of two folded C-rich strands. The discovery of i-motif structures in the two types of alpha-satellite monomers leads us to suggest a potential role of this non-canonical structure in centromeric chromatin organization.

Materials and Methods

Sample preparation

Oligonucleotides were purchased from Integrated DNA Technologies, IDT, Coralville, IA, USA. Samples for NMR experiments were dissolved in either D_2O or 9:1 H_2O/D_2O. Experiments were performed either with no buffer added or in 25 mM sodium phosphate buffer, and 100 mM NaCl. Experiments were carried out at different pHs, ranging from 3.5 to 7. pH was adjusted by adding aliquots of either concentrated solution of DCl or NaOD.

NMR experiments

All NMR spectra were acquired in Bruker spectrometers operating at 600 and 800 MHz, equipped with cryoprobes and processed with the TOPSPIN software. In the experiments in D_2O, presaturation was used to suppress the residual H_2O signal. A jump-and-return pulse sequence[31] was employed to observe the rapidly exchanging protons in 1D H_2O experiments. NOESY spectra in D_2O and 9:1 H_2O/D_2O were acquired with mixing times of 50, 150, 250 and 300 ms. TOCSY spectra were recorded with the standard MLEV-17 spin-lock sequence and a mixing time of 80 ms. In most of the experiments in H_2O, water suppression was achieved by including a WATERGATE module in the pulse sequence prior to acquisition. The spectral analysis program SPARKY[32] was used for semiautomatic assignment of the NOESY cross-peaks and quantitative evaluation of the NOE intensities.

Circular Dichroism spectroscopy

Circular dichroism spectra at different temperatures were recorded on a Jasco J-810 spectropolarimeter fitted with a thermostated cell holder. CD spectra were recorded in 25 mM

sodium phosphate buffer, pH 4, with 100 mM NaCl (100 μM oligo concentration). CD melting curves were recorded at the wavelength of the larger positive band, 285 nm, with a heating rate of 0.5 ºC·min^{-1}. Titration experiments were performed by recording the absorbance at λ = 285 nm and different pH values. pH was adjusted by adding aliquots of concentrated solutions of DCl or NaOD.

Mass spectrometry

All ESI-MS experiments were carried out in the negative ion mode on an Exactive ESI-Orbitrap mass spectrometer (Thermo Scientific, Bremen, Germany). The ESI spray voltage and capillary voltage used were -2.75 kV and -20 V, respectively. The capillary temperature was set to 150 ºC. Tube lens and skimmer voltage were fixed to 180 V and -10 V, respectively. All the oligonucleotides analysed were dissolved at 100 μM in 100 mM NH_4OAc buffer at pH 7 and pH 4, and were injected at a flow rate of 4 μL min^{-1}.

NMR constraints

Initial calculations were performed with qualitative distance constraints (classified as 3, 4 or 5 Å) and the resulting structures were then refined by employing more accurate distance constraints obtained from a complete relaxation matrix analysis with the program MARDIGRAS[33]. Error bounds in the interprotonic distances were estimated by carrying out several MARDIGRAS calculations with different initial models, mixing times and correlation times, as described in previous works. In addition to these experimentally derived constraints, hydrogen bond restraints were used. Target values for distances and angles related to hydrogen bonds were set to values obtained from crystallographic data in related structures[34].

Structure determination

Structures were calculated with the program DYANA 1.4[35] and further refined with the SANDER module of the molecular dynamics package AMBER 7.0[36]. Initial DYANA calculations were carried out on the basis of qualitative distance constraints. The resulting structures were used as initial models in the complete relaxation matrix calculations to obtain accurate distance constraints, as described in the previous paragraph. These structures were taken as starting points for the AMBER refinement, consisting of an annealing protocol in vacuo, followed by long trajectories where explicit solvent molecules were included and using the Particle Mesh Ewald method to evaluate long-range electrostatic interactions. The specific protocols for these calculations have been describe elsewhere[37]. The AMBER-98 force field[38] was used to describe the DNA, and the TIP3P model was used to simulate water molecules. Analysis of the representative structures as well as the MD trajectories was carried out with the programs Curves V5.1[39] and MOLMOL[40].

Results

The A box of the human alpha-satellite DNA contains tracts of two-cytidines.

The consensus sequence of the human A box is 5'-TTCCTTTTCPyACCATAG-3'[13] where Py can be C (**HS**) or T (**HST**) in the autosomes and the X chromosome, but it is mainly C in the Y chromosome. In this study we have also used two truncated versions of **HS** (**HS1** and **HS2)** (Figure 2). The residue numbering used in this paper is shown in Figure 2a, and it has been defined arbitrarily for convenient discussion of the NMR spectra and structures.

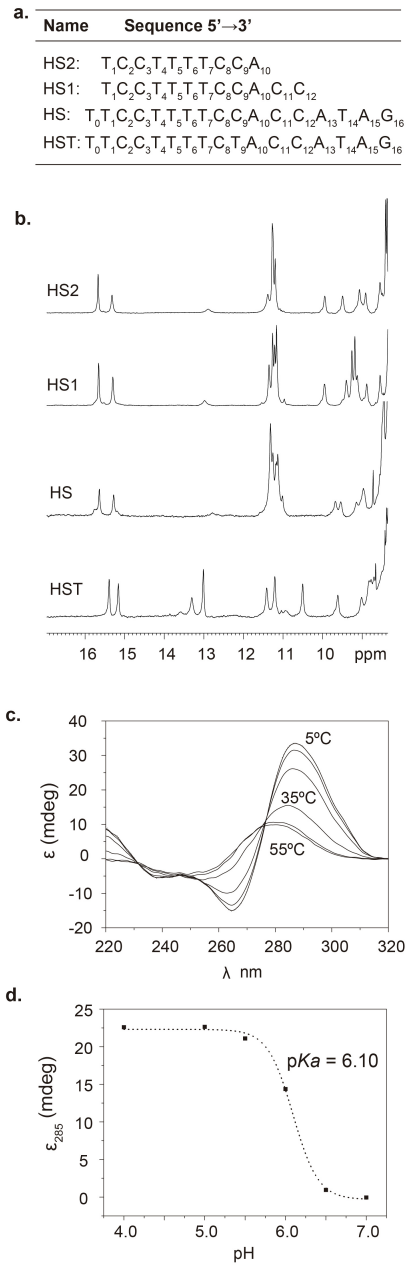


***Figure 2**. a) Sequences of the oligonucleotides studied. b) Exchangeable protons region of the NMR spectra of the centromeric sequences in 25 mM NaPi, 100 mM NaCl buffer at pH 4.0 (HS, HS1 and HS2) and pH 4.5 (HST) T=5 ºC (Oligo concentrations: [HS2] = 2.0 mM, [HS1] = 2.0 mM, [HS] = 0.8 mM and [HST] = 0.8 mM). c) CD spectra of **HS2** at different temperatures (Buffer conditions: 25 mM NaPi, 100 mM NaCl pH=4) ([HS2] = 100 µM). d) pH titration of HS2 at T=5ºC ([HS2] = 100 µM).*

NMR, CD and mass spectrometry show that the C-rich strand of the A boxes form dimeric i-motifs.

Mass spectra of **HS1** and **HS2** at different pH conditions are shown in Supplementary Figure S1. In all cases, dimeric species are clearly detected, and they are more abundant at acidic pH. The isotopic distributions shown in Supplementary Figure S1 are strikingly identical at the two pH values, showing a separation between consecutive ^{13}C isotopes consistent with the presence of a dimer (see figure legend). The CD spectra at acidic pH exhibit a positive band at around 285 nm and a negative one around 265 nm, which are characteristic of i-motif structures[41] (Figure 2c). Both bands disappear when the temperature or pH increases, indicating a pH dependent denaturation process. The midpoint pH of this transition provides an apparent pK_a value for the overall structure of **HS2** of 6.1 (Figure 2d). Melting transitions were monitored by CD obtaining Tm values of 26.6 and 30.5 ºC for **HS1** and **HS2,** respectively (Supplementary Figure S2).

Under acidic conditions, proton NMR spectra show signals characteristic of non-canonical base pairs (Figure 2b and Supplementary Figure S2). The signals around 15.4 ppm are distinctive of cytosine imino protons involved in hemi-protonated C:C$^+$ base pairs (as well as the signals at around 9.0-10.0 ppm for their amino protons)[21,22]. The sharp signals between 10.0 and 11.5 ppm may correspond to other non-canonical base pairs, most probably wobble T:T base pairs. The signal around 12.7 ppm corresponds to an A:T base pair. This signal becomes broad at pH 4.0 and disappears at pH 4.5 (see Supplementary Figure S3). Except by this broadening effect, the general features of the NMR spectra are very similar in a wide range of pH (3.5 to 6.5), in sodium and ammonium acetate buffer (the latter required for mass spectrometry experiments), and in a range of oligonucleotide concentrations between 100 μM and 3 mM (Supplementary Figure S1 and S4). This indicates that the species observed in the NMR and CD experiments are the same dimeric species observed in the mass spectrometry experiments.

Assignment of the NMR spectra and general structural features.

The NMR spectra of the oligonucleotides **HS**, containing the main variant of the full human A box, together with its truncated versions (**HS1** and **HS2**), are shown in Figure 2b. In the three cases, NMR spectra are very similar. This resemblance is also observed in the two-dimensional spectra (see Figure 3a and Supplementary Figures S5-6), where almost identical NOE patterns between exchangeable protons are observed. This indicates that the three constructions are dimers with the same symmetry and stabilized by the same base pairs (number of C:C$^+$, etc.), with only residues 1 to 10 involved in base pair formation. This strong similarity allowed us to focus in the NMR spectra of **HS1** and **HS2** for resonance assignments.

Sequential assignment of exchangeable and non-exchangeable protons was conducted following standard methods (assignment lists are shown in the Supplementary Material). The number of major signals found in the NMR spectra of **HS2** and **HS1** were 10 and 12, respectively,

indicating the two subunits in the dimeric structures are magnetically equivalent. Additional broad and low intensity signals are observed in the spectra recorded at 5 ºC. These signals become sharper and more intense at higher temperature, becoming majoritary at T > 25-30 ºC, which indicates that correspond with the unfolded species. This behavior, illustrated in Figure S7 for the cytosine H5-H6 cross-peaks, is characteristic of equilibrium between the folded and unfolded species, which is slow in the NMR time scale.

Sequential contacts with A10 and T1 allowed the assignment of C9 and C2. The stacking order could be followed by H1'-H1' connectivities between C2-C8, C2-C9 and C3-T7 (Figure 3c), and other NOEs such as H6C9-H1'C2, H6C3-H1'T7. Sequential sugar-aromatic cross-peaks were found between thymine loop residues, allowing the assignment of T5, T6 and T7. The remaining thymine spin system did not present sequential connectivities and it was assigned to T4. In the case of **HS1**, some sequential NOEs were observed for residues A10→C11 and to a less extension for C11→C12. No sequential NOEs were observed for residues 13 to 16 in **HS**, suggesting that these residues are disordered in solution. According to the H6/8-H1' NOE intensities all the glycosidic angles are anti.

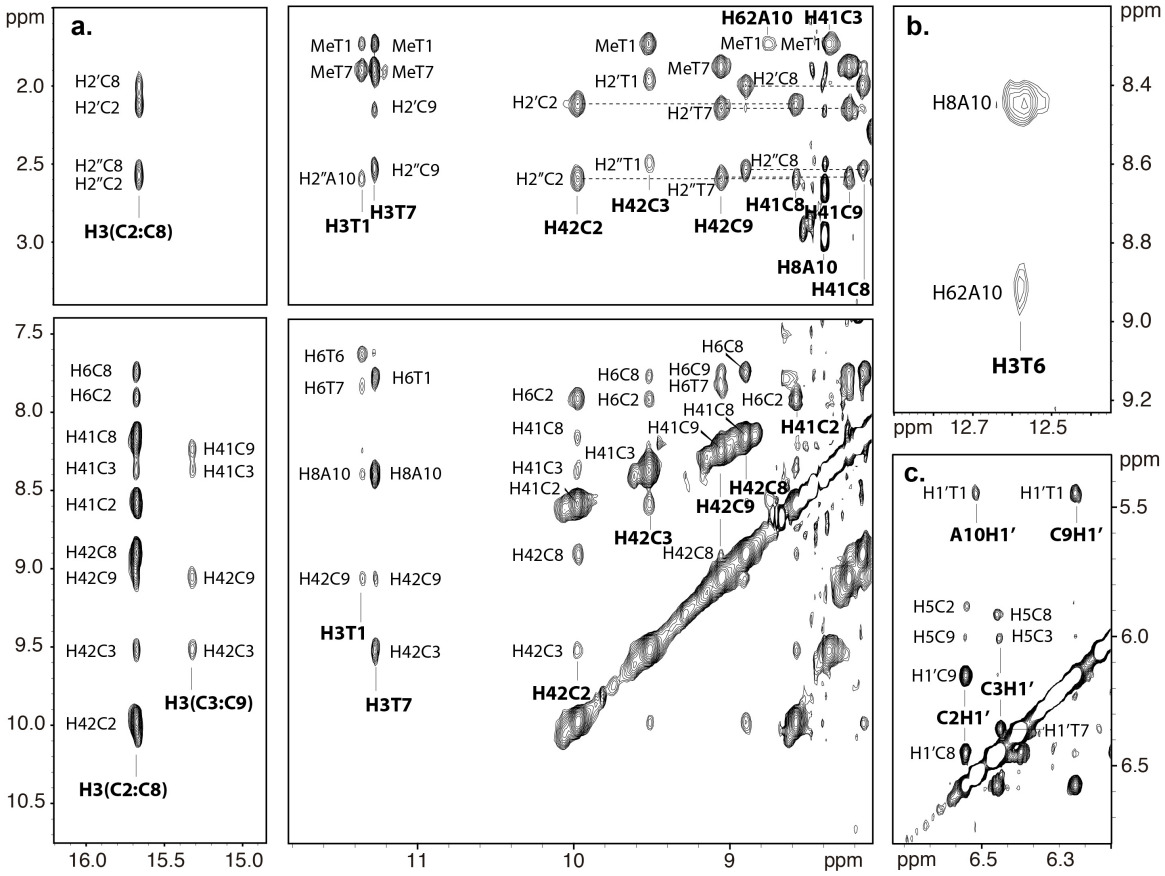


Figure 3. *Regions of NOESY spectra of **HS2** at pH 3.6 and T=5 ºC ([HS2] = 2.9 mM)**. a)** Four panels with exchangeable proton regions of NOESY spectra of **HS2** in 90/10 H_2O/D_2O (mixing time 150 ms). b) Detail of NOESY spectra of **HS2** in 90/10 H_2O/D_2O (mixing time 50 ms) showing a Hoogsteen NOE pattern. C) Region of the NOESY spectra of **HS2** in D_2O, showing characteristic H1'-H1' correlations.*

The exchangeable proton spectra exhibit two signals at 15.68 and 15.32 ppm, indicative of hemiprotonated C:C^+ base pairs (Figure 2). Each of these signals presents four cross-peaks with amino protons in the 8-10 ppm range (Figure 3a). This indicates that the two C:C^+ base pairs occur between non magnetically equivalent cytosines. NOEs between amino protons of C2 and C8 with H2'/H2" of the same residue are clearly observed (Figure 3a). Since these cross-peaks are characteristic of i-motifs and they occur through the major groove formed by antiparallel oriented neighbor strands, the only possible orientation between both subunits is head-to-tail, with the minor grooves occurring between intra-molecular strands. This orientation is confirmed by the H1'-H1' contacts mentioned above, and amino-H2'H2" contacts between C9 and T7, and C3 and T1 (Figure 3a). All the exchangeable and non-exchangeable NOEs are consistent with the following stacking order: A10-T1-C9-C2-C8-C3-T7. Additional NOEs between amino protons of C9 and methyl of T7, and amino protons of C3 and of methyl T1 confirm that C3-C9 base pair are located at the end of the cytosine stack. This location is also supported by the exchangeable protons of C3 and C9, which exhibit broader lines and disappear at lower temperatures than those of C2 and C8 (Figure S2).

Several thymine imino protons are found between 10.5 and 11.5 ppm. The signals at 11.36 and 11.17 exhibit strong cross-peaks between them, indicating the formation of a T:T base pair. A number of contacts with the neighboring residues allow the assignment of these signals to the imino protons of T1 and T7, respectively (Figure 3a). Cross-peaks with C3 and C9 protons indicate that T1:T7 base pair is directly in contact with the C3:C9 base pair. At pH 3.6, an intense cross-peak between H8A10 with a thymine imino proton is observed, indicating the formation of a Hoogsteen base pair (Figure 3b). This imino proton was assigned to T6, since the assignment to T4 or T5 gave rise to strong constraint violations during the structural calculations. Moreover, a number of NOEs between protons of T1 and T6 with T7 and A10 were clearly detected, indicating that the T6:A10 Hoogsteen base pair is on top of T1:T7 (Figure 3). At higher pH, the signal of H3T6 becomes broader and disappears at pH 4.5. The only other signals affected in this range of pH are H2 and H8 of A10. The effect of pH in their chemical shifts is shown in Figure S3. The pH titration curves exhibit a sigmoidal dependence with a midpoint value of 4.1, consistent with reported pK_a values for adenine protonation in position N1. This apparent pK_a value is slightly higher than in the free mononucleotide (pK_a = 3.5).[42,43]

As mentioned above, the exchangeable proton spectra of **HS1** indicate that C11 and C12 are not involved in the core of intercalated cytosine pairs. Sequential sugar-base NOEs suggest that C11 stacks on top of A10. The lack of sequential NOEs for the remaining residues at the 3'-end of **HS**, and the fact that no exchangeable protons are observed for residues 13 to 16, indicate that these residues are mainly disordered.

Sugar conformation can be deduced from qualitative analysis of DQF-COSY spectra. H1'-H2'/2" cross-peaks for residues T4 to T7, and A10 are consistent with J1'2' and J1'2" coupling

constants in the range of 7-10 and 6-7 Hz, respectively. This indicates that the sugar conformations of these residues are in the general South domain. However, residues T1 and C2 exhibit very small J1'2', indicating that these deoxyriboses adopt a North conformation. No direct experimental evidence of the sugar conformation of C3, C8 and C9 was obtained due to signal overlap.

Solution structure of HS2.

The three dimensional structure of **HS2** was calculated on the basis of 300 experimental distance constraints by using restrained molecular dynamics methods, and following standard procedures used in our group. Except for T4 and the corresponding residue in the symmetry related sub-unit, all residues are well defined, with an RMSD of 0.8 Å (Table S4). The final AMBER energies and NOE terms are reasonably low in all the structures, with no average distance constraint violation>0.7 Å.

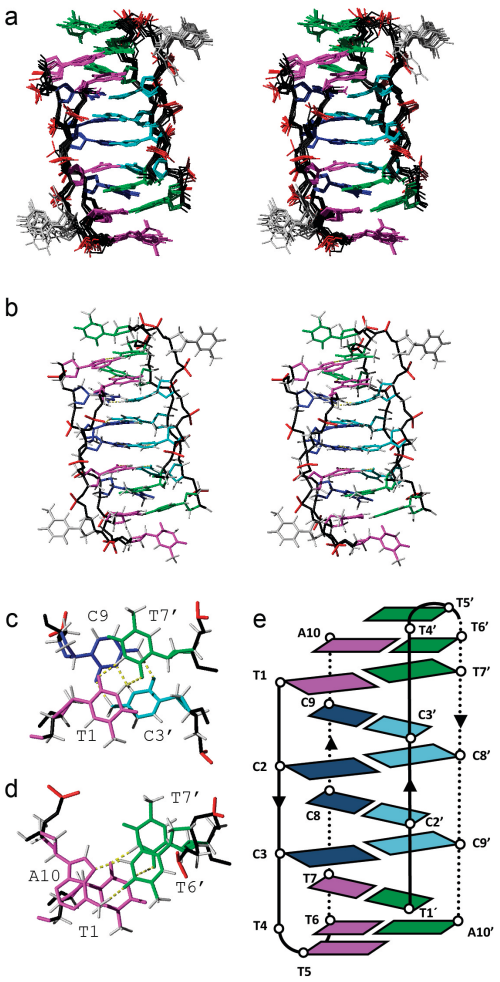


***Figure 4**. Dimeric structure of **HS2**. a) Stereoview of the ensemble of the 10 calculated structures. b) Stereoview of the average structure. c) Detail of the stacking interaction between C:C$^+$ and T:T base pairs. d) Detail of the stacking interaction between T:T and the capping Hoogsteen A:T base pair. e). Schematic representation of the dimeric structure of **HS2**. Colour code: Cytosines in the two sub-units are shown in blue and cyan, respectively; well-defined thymines and adenines in the two sub-units are shown in magenta and green, respectively; no well-defined residues in grey; and backbone in black. Hydrogen bonds are indicated in yellow.*

The resulting structure is a dimer, consisting of two molecules of d(TCCTTTTCCA) arranged head-to-tail (see schematic representation in Figure 4e). As reflected by the number of signals in the NMR spectra, the dimer is symmetric. The two decamers associate with each other by forming four intercalated hemi-protonated $C:C^+$ base pairs (C2-C8 and C3-C9, and their symmetry related counterparts), sandwiched by two intermolecular T1:T7 base pairs. The structure is capped by two intermolecular T6:A10 Hoogsteen base pairs, and the unpaired T5. The base-paired cytidines present the characteristic sugar-sugar contacts between adjacent strands through the minor groove. The two sides of $C:C^+$ stacks correspond with the cytosine located at 3'-end of the C-tracts (3'E type of i-motif structures [28-30]). The two T:T base pairs, contiguous to the $C:C^+$ core stack, follow the same pattern of alternate base pair between parallel oriented strands (Figure 4c). This feature confirms the ability of thymines to fit into intercalative $C:C^+$ structures[44]. However, the Hoogsteen AT base pairs occur between antiparallel oriented strands and do not follow the alternate base pair motif (Figure 4d). In spite of this, extensive stacking interactions occur between T6:A10 and T1:T7 base pairs. As shown in Figure 4a, residue T4 and its symmetry related one are mainly disordered.

All glycosidic angles are *anti*, with values ranging from -110º to -150º, with the only exception of cytosines at the 3'-end of the tracts (C8, C9 and their symmetric counterparts) which adopt a high *syn* glycosidic conformation of around -70º to -90º. Sugar pucker of T1, C2, C3 and C9 are in the North domain, C8 adopts an East conformation, and the remaining residues are in the general South domain. The dihedral torsion angles are shown in Table S5.

Solution structure of the human A box (HS)

As mentioned above, all the NMR evidences indicate that the structure of **HS1** and **HS** are very similar to that of **HS2**. In fact, most signals in **HS1** and **HS2** spectra are almost identical (see Figures 3, and S5-6). Only some weak sequential cross-peaks between A10→C11→C12 were detected in **HS1**'s NOESY spectra, suggesting that C11 and C12 may be relatively ordered, although they are not involved in additional base pairs. This is consistent with the CD melting experiments, which show that addition of these two residues does not confer any extra stability to **HS1** vs. **HS2** (Supplementary Figure S2). In the case of **HS**, the NMR spectra indicate that the 3'-terminal tail (residues A13 to G16) is completely disordered. The small signals observed in the cytosine imino regions at low temperature suggest the formation of an additional species. According to the signal intensity this minor species has a very low population of less than 10% and it is only observed a low temperature. Taking together the NMR spectra and the mass spectrometry data mentioned above, we can conclude that the major structure of the complete human A box is the dimer represented in Figure 5b. The structure consists of two well-defined loops (residues 1 to 10) that interact with each other through the formation of four intercalative $C:C^+$, two T:T and two A:T Hoogsteen base pairs with the remaining residues (11 to 16) being mainly disordered.

The structure of the main variant of the human A box (C9→T9).

As shown in Supplementary Figure S9a, mass spectrometry data show that the oligonucleotide **HST** forms dimeric structures at neutral pH. The NMR spectra of **HST**, shown in Figure 2b and S9b, exhibit the characteristic signals of $C:C^+$ base pairs. The analysis of the 2D spectra clearly shows that the dimer is symmetrical (the two subunits are equivalent) and is stabilized by the formation of four $C:C^+$ base pairs. Two imino signals at 15.39 and 15.14 ppm, corresponding to the hemiprotonated $C:C^+$ base pairs, exhibit four cross-peaks with cytosine amino protons, indicating the $C:C^+$ base pairs occur between non-equivalent residues. Spectral assignment of the cytosine core can be carried out by standard methods and shows that these pairs are C2-C11 and C3-C12 (see Figure 5). The only possible arrangement of a symmetric dimer with these base pairs is a head-to-tail association of the two subunits, with the stacking order T1-A13-C2-C12-C3-C11-T4-A10. This order could be confirmed by a number of inter-strand H1'-H1' cross-peaks along the minor groove, and inter-strand amino-H2'H2" cross-peaks through the major groove. This indicates that the core of $C:C^+$ base pairs in the dimeric structures of **HS** and **HST** are very similar, with the role of cytosines C8 and C9 in the main variant (**HS**) being substituted by C11 and C12 in the C9→T9 variant (**HST)** (see Figure 5c and S10).

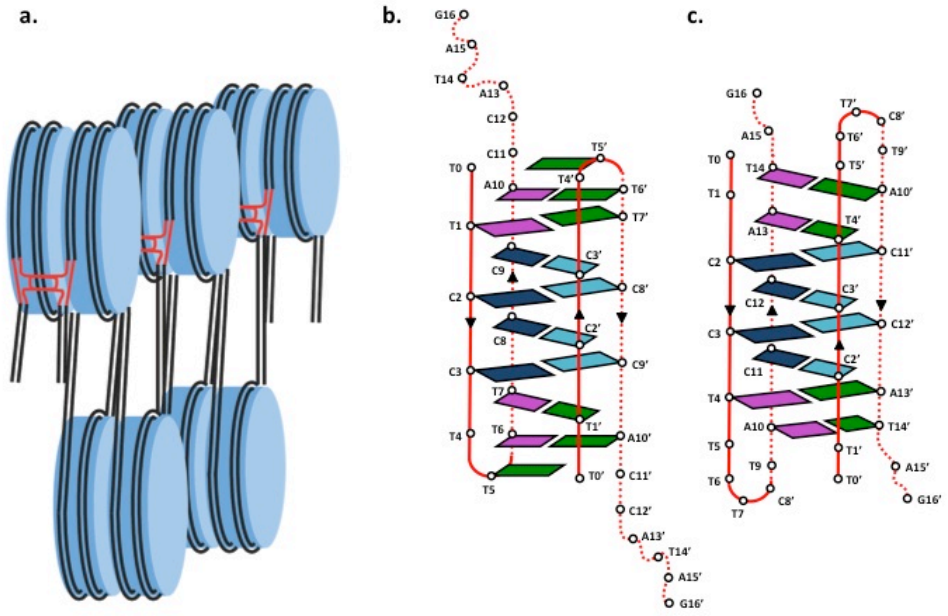


***Figure 5**. a) Model for nucleosome bundling in centromeric chromatin. B/A boxes of the alpha–satellite DNA (in red) would maintain the structural organization of the centromere by forming dimeric i-motif structures, like those found in vitro for the two main variants of the alpha-satellite A box, shown schematically in panels b), **HS,** and c) **HST**.*

In spite of these similarities in the core, the NMR spectra indicate that there are significant differences in the loop regions. Of particular relevance are the two signals around 13 ppm, indicative of formation of two A:T base pairs, which corresponds with four base pairs in the dimer.

The NOE cross-peaks between the imino at 13.08 ppm and adenine H2 and amino protons indicate the formation of a Watson-Crick or reverse Watson-Crick base pair. A number of NOEs with the neighbor C2:C11 base pair suggest that this A:T pair is intermolecular and corresponds with T4:A13 (Figure 5c and S10). The imino signal at 12.95 ppm exhibits a strong cross-peak with the H8A10, indicating the formation of a Hoogsteen base pair. With the experimental data available, this imino resonance could be assigned to T1 or T14, giving rise to the base pairs T1:A10 or T14:A10, respectively. Model calculations with these two possibilities and the experimental distance constraints derived from unambiguously assigned NOEs indicate that a Hoogsteen T14:A10 base pair gives rise to better structures (no significant distortions, and lower distance constraints violations). Although a complete structural determination is necessary to provide a full picture of the structure of the loops in **HST**, we can conclude that the structure is a head-to-tail association of two subunits stabilized by four hemiprotonated $C:C^+$ base pairs and four A:T base pairs, as shown schematically in Figure 5c.

Discussion

The structures of the A box and B box of the human alpha-satellite are examples of self-recognition in DNA sequences. In a previous study, Gallego et al.[18,19] show that the B box forms an i-motif resulting from the dimerization of two folded C-rich strands in a head-to-head way. The dimer is stabilized by a tract of five $C:C^+$ hemiprotonated base pairs capped at one end by a G:T:G:T minor groove tetrad and an A:A base pair at the other side. On the other hand, we show here that the A box also forms a dimeric i-motif. In this case, the two folded C-rich strands associate in a head-to-tail way and interact through formation of four $C:C^+$ base pairs, and other additional interactions that differ in each of the two main A box variants. In the main variant, the $C:C^+$ core is extended at each side by T:T base pairs, which follow the same pattern of intercalated base pairs between parallel oriented strands. The structure is capped at both ends by Hoogsteen A:T base pairs between antiparallel oriented strands and an unpaired thymine. In the second variant, the C9→T9 mutation provokes a rearrangement of the i-motif core, which consists now of C2-C3 and C11-C12 tracts. Consequently the loops in **HST** structure are much larger than in **HS** (Figure 5). Stacking order between $C:C^+$ base pairs differs between both structures, being 3'E in **HS** and 5'E in **HST**. The closing T:T base pairs in **HS** are now substituted by A:T base pairs, most probably in a reverse Watson-Crick conformation. This difference is consistent with a recent report by Lieblein et al., showing that adenines next to a $C:C^+$ base pairs favor 5'E conformations in i-motifs[45]. Interestingly, both structures are capped by Hoogsteen A:T base pairs between antiparallel oriented strands.

T:T base pairs are well accommodated in i-motif structures[46] and they have been observed in loops connecting C-tracts[47]. A:T base pairs are not so common, but they have been also observed in monomeric[20,47,48] and tetrameric i-motifs[49]. In some cases, adenine protonation affects

the structure[20,50] and the stability of the i-motif. Generally, this effect is destabilizing since protonation at N1 position disrupts Watson-Crick base pairs[20,50]. In our case, adenine protonation in N1 results in stabilization of an A:T Hoogsteen base pair, probably due to acidification of the adenine amino protons, enabling them to form stronger hydrogen bonds. This stabilization of Hoogsteen base pair upon adenine protonation has been reported in some RNA molecules, like the hairpin ribozyme[43].

As usual in i-motif structures, the $C:C^+$ pairing provokes the formation of four grooves; two major grooves and two extremely narrow minor grooves. Interestingly, in all these structures the minor grooves are intra-molecular, as reported in other dimeric i-motif structures[44,46,51]. This is in contrast with the dimeric i-motif formed by d(5mCCTTTACC)[52] where the minor groove is inter-molecular.

All these dimeric structures show that $C:C^+$ intercalative base pairs constitute a robust motif for self-recognition between DNA sequences. Interestingly, the structures of **HS** and **HST** obtained in this study exemplify how the sequences connecting short C-tracts are able to undergo large structural re-adjustments to maintain the intercalated $C:C^+$ core and preserve dimerization. The ability of diverse (and apparently unrelated) sequences connecting C-tracts to form stable i-motifs makes the identification of these dimeric motifs at primary sequence level very difficult.

Recent *in vivo* evidence for phasing of humans and rice CENP-A nucleosomes on centromeric satellites has shown that nucleosome positioning is a physical requirement for centromere formation[5,53]. Moreover, human, mouse, rice and *Drosophila* centromeres contain blocks of CENP-A nucleosomes that are interspersed with blocks of canonical histone H3 nucleosomes[2,54], and although the folding of this centromeric chromatin is still unresolved, it is assumed that CENP-A nucleosomes self-associate laterally and exclude histone H3 containing nucleosomes to form a lattice on the surface of the chromosome primary constriction. The flexibility observed in the DNA regions located at the entrance and the exit of CENP-A nucleosomes may be the physical feature that facilitates the lateral interactions[5,6].

In humans, the phasing of CENP-A nucleosomes on alpha-satellite DNA place the A box and the CENP-B box at the entrance and exit of the nucleosome[5]. In mice, the CENP-B box also appears at the end of nucleosome core particles[55]. Moreover, the binding of CENP-B to CENP-A nucleosomes does not protect the CENP-B box from nuclease digestion[56]. Therefore, these regions remain accessible in each CENP-A nucleosome.

The function of mammalian CENP-Bs in natural centromeres still remains mysterious. On the one hand, higher-order alpha-satellite repeats containing CENP-B boxes and the CENP-B itself appear to be required for *de novo* centromere formation in human artificial chromosome assays[1,57-60], and it has been shown that CENP-B provides a parallel pathway for kinetochore formation[61]. On the other hand, CENP-B is not an essential centromeric protein because it is absent from functional centromeres that lack CENP-B boxes (human and mouse Y centromeres and human

neocentromeres). Moreover, CENP-B null mice are viable and do not show mitotic or meiotic defects[62-64].

Recently, using both CENP-A chromatin immunoprecipitation followed by sequencing analysis and artificial chromosome assays, Willard's group has been able to demonstrate that it is not just the presence of CENP-B boxes but rather the pattern of CENP-B boxes and A boxes within the higher-order repeat array that contributes to *de novo* centromere formation[1]. Therefore, a hierarchical mechanism of chromatin folding based on CENP-B boxes and A boxes interactions may determine the three-dimensional structure of the centromere.

The discovery of i-motif structures in the two types of alpha-satellites (A box and B box) supports the idea that this non-canonical DNA structure may have a role in the structural organization of the centromere. The biochemical data discussed above together with our structural findings are consistent with the centromeric nucleosome organization shown in Figure 5. Thus, it is expected that the higher stability of CENP-B box or A box i-motif homo-dimers versus CENP-B box/A box i-motif hetero-dimers would prevent out-of-register interaction of higher-order repeat units and, in turn, would determine the ordered spatial arrangement of the centromeric chromatin in metaphase chromosomes. Moreover, the high stability of the i-motifs may confer an enhanced resistance against the pulling forces felt by the centromeric chromatin during chromosome segregation. In this sense, it is interesting to note the recent report by Famulok et al, showing how DNA nanocircles containing C-tracts can be arranged in well-defined assemblies through formation of intermolecular i-motifs[65]. Likewise, I-motifs may induce the assembly of more complex multicomponent DNA architectures in the cell.

All together, these results support our initial hypothesis that centromeric alpha-satellite DNA may have been selected not by their primary sequence but by their ability to form i-motif structures. If this were the case, the "centromere paradox" may well be explained by shared secondary structures without shared primary structures.

Coordinates

Atomic coordinates of **HS2**, d(TCCTTTTCCA), have been deposited in the Protein Data Bank (accession number 2MRZ).

Acknowledgements

We gratefully acknowledge Dr. Douglas V. Laurents for revision of the manuscript and his useful comments. We also thank the Structural Biophysical Chemistry platform of the IECB (CNRS UMS3033 / Inserm US001) for the access to the mass spectrometry facility and Dr. Frederic Rosu for his kind assistance.

Funding

This work was supported by the MICINN (CTQ2010-21567-C02-02 to CG, BFU2011-30295-C02-01 to AV), the Inserm (ATIP-Avenir Grant no. R12086GS to V.G.), the Conseil Régional Aquitaine (Grant no. 20121304005 to V.G.), the EU (FP7-PEOPLE-2012-CIG-333611 to V.G.), and the institutional grant from the Fundación Ramón Areces to the Centro de Biología Molecular "Severo Ochoa"). MG was supported by the FPI-fellowship BES-2009-027909.

References

1 Hayden, K. E. *et al.* Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **33**, 763-772 (2013).

2 Allshire, R. C. & Karpen, G. H. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nat. Rev. Genet.* **9**, 923-937 (2008).

3 Black, B. E. & Cleveland, D. W. Epigenetic centromere propagation and the nature of CENP-a nucleosomes. *Cell* **144**, 471-479 (2011).

4 Fachinetti, D. *et al.* A two-step mechanism for epigenetic specification of centromere identity and function. *Nat. Cell Biol.* **15**, 1056-1066 (2013).

5 Hasson, D. *et al.* The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nat. Struct. Mol. Biol.* **20**, 687-695 (2013).

6 Panchenko, T. *et al.* Replacement of histone H3 with CENP-A directs global nucleosome array condensation and loosening of nucleosome superhelical termini. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16588-16593 (2011).

7 Malik, H. S. & Henikoff, S. Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. *Genetics* **157**, 1293-1298 (2001).

8 Talbert, P. B., Bryson, T. D. & Henikoff, S. Adaptive evolution of centromere proteins in plants and animals. *J. Biol.* **3**, 18 (2004).

9 Schueler, M. G., Swanson, W., Thomas, P. J. & Green, E. D. Adaptive evolution of foundation kinetochore proteins in primates. *Mol. Biol. Evol.* **27**, 1585-1597 (2010).

10 Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098-1102 (2001).

11 Abad, J. P. & Villasante, A. Searching for a common centromeric structural motif: *Drosophila* centromeric satellite DNAs show propensity to form telomeric-like unusual DNA structures. *Genetica* **109**, 71-75 (2000).

12 Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. Alpha-satellite DNA of primates: old and new families. *Chromosoma* **110**, 253-266 (2001).

13 Romanova, L. *et al.* Evidence for selection in evolution of alpha satellite DNA: the central role of CENP-B/pJα binding region. *J. Mol. Biol.* **261**, 334-340 (1996).

14 Tyler-Smith, C. & Brown, W. R. Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.* **195**, 457-470 (1987).

15 Kalitsis, P., Griffiths, B. & Choo, K. H. Mouse telocentric sequences reveal a high rate of homogenization and possible role in Robertsonian translocation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8786-8791 (2006).

16 Pertile, M. D., Graham, A. N., Choo, K. A. & Kalitsis, P. Rapid evolution of mouse Y centromere repeat DNA belies recent sequence stability. *Genome Res.* **19**, 2202-2213 (2009).

17 Yoda, K., Kitagawa, K., Masumoto, H., Muro, Y. & Okazaki, T. A human centromere protein, CENP-B, has a DNA binding domain containing four potential alpha helices at the NH2 terminus, which is separable from dimerizing activity. *J. Cell Biol.* **119**, 1413-1427 (1992).

18 Gallego, J., Chou, S.-H. & Reid, B. R. Centromeric pyrimidine strands fold into an intercalated motif by forming a double hairpin with a Novel T:G:G:T tetrad: solution structure of the d(TCCCGTTTCCA) dimer. *J. Mol. Biol.* **273**, 840-856 (1997).

19 Gallego, J., Golden, E. B., Stanley, D. E. & Reid, B. R. The folding of centromeric DNA strands into intercalated structures: a physicochemical and computational study. *J. Mol. Biol.* **285**, 1039-1052 (1999).

20 Nonin-Lecomte, S. & Leroy, J. L. Structure of a C-rich strand fragment of the human centromeric satellite III: a pH-dependent intercalation topology. *J. Mol. Biol.* **309**, 491-506 (2001).

21 Gehring, K., Leroy, J.-L. & Gueron, M. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature* **363**, 561-565 (1993).

22 Leroy, J. L., Gehring, K., Kettani, A. & Gueron, M. Acid multimers of oligodeoxycytidine strands: stoichiometry, base-pair characterization, and proton exchange properties. *Biochemistry* **32**, 6019-6031 (1993).

23 Lieblein, A. L., Kramer, M., Dreuw, A., Furtig, B. & Schwalbe, H. The nature of hydrogen bonds in cytidine...H+...cytidine DNA base pairs. *Angew. Chem. Int. Ed. Engl.* **51**, 4067-4070 (2012).

24 Guo, K. *et al.* Formation of pseudosymmetrical G-quadruplex and i-motif structures in the proximal promoter region of the RET oncogene. *J. Am. Chem. Soc.* **129**, 10220-10228 (2007).

25 Cui, J., Waltman, P., Le, V. H. & Lewis, E. A. The effect of molecular crowding on the stability of human c-MYC promoter sequence i-motif at neutral pH. *Molecules* **18**, 12751-12767 (2013).

26 Sun, D. & Hurley, L. H. The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression. *J. Med. Chem.* **52**, 2863-2874 (2009).

27 Chan, F. L. *et al.* Active transcription and essential role of RNA polymerase II at the centromere during mitosis. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1979-1984 (2012).

28 Benabou, S., Avino, A., Eritja, R., Gonzalez, C. & Gargallo, R. Fundamental aspects of the nucleic acid i-motif structures. *RSC Advances* **4**, 26956-26980 (2014).

29 Dong, Y., Yang, Z. & Liu, D. DNA Nanotechnology Based on i-Motif Structures. *Acc. Chem. Res.* **47**, 1853-1860 (2014).

30 Day, H. A., Pavlou, P. & Waller, Z. A. E. i-Motif DNA: Structure, stability and targeting with ligands. *Bioorg. Med. Chem.* **22**, 4407-4418 (2014).

31 Plateau, P. & Gueron, M. Exchangeable proton NMR without base-line distorsion, using new strong-pulse sequences. *J. Am. Chem. Soc.* **104**, 7310-7311 (1982).

32 SPARKY v. 3. (University of California, San Francisco).

33 Borgias, B. A. & James, T. L. MARDIGRAS, a procedure for atrix analysis of relaxation for discerning geometry of an aqueous structure. *J. Magn. Reson.* **87**, 475-487 (1990).

34 Cai, L. *et al.* Intercalated cytosine motif and novel adenine clusters in the crystal structure of the *Tetrahymena* telomere. *Nucleic Acids Res.* **26**, 4696-4705 (1998).

35 Guntert, P., Mumenthaler, C. & Wuthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283-298 (1997).

36 Case, D. A. *et al. AMBER 7*, 2002).

37 Soliva, R. *et al.* Solution structure of a DNA duplex with a chiral alkyl phosphonate moiety. *Nucleic Acids Res.* **29**, 2973-2985 (2001).

38 Cornell, W. D. *et al.* A 2nd generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **117**, 5179-5197 (1995).

39 CURVES, helical analysis of irregular nucleic acids. v. 3.0 (Laboratory of Theoretical Biochemistry CNRS, Paris, 1990).

40 Koradi, R., Billeter, M. & Wuthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14**, 29-32 (1996).

41 Kypr, J., Kejnovská, I., Renčiuk, D. & Vorlíčková, M. Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res.* **37**, 1713-1725 (2009).

42 Guéron, M. & Leroy, J.-L. [16] Studies of base pair kinetics by NMR measurement of proton exchange. *Methods Enzymol.* **261**, 383-413 (1995).

43 Ravindranathan, S., Butcher, S. E. & Feigon, J. Adenine protonation in domain B of the hairpin ribozyme. *Biochemistry* **39**, 16026-16032 (2000).

44 Canalia, M. & Leroy, J. L. Structure, internal motions and association–dissociation kinetics of the i-motif dimer of d (5mCCTCACTCC). *Nucleic Acids Res.* **33**, 5471-5481 (2005).

45 Lieblein, A. L., Fürtig, B. & Schwalbe, H. Optimizing the Kinetics and Thermodynamics of DNA i-Motif Folding. *ChemBioChem* **14**, 1226-1230 (2013).

46 Canalia, M. & Leroy, J. L. [5mCCTCTCTCC]4: an i-motif tetramer with intercalated T*T pairs. *J. Am. Chem. Soc.* **131**, 12870-12871 (2009).

47 Han, X., Leroy, J.-L. & Guéron, M. An intramolecular i-motif: the solution structure and base-pair opening kinetics of d(5mCCT3CCT3ACCT3CC). *J. Mol. Biol.* **278**, 949-965 (1998).

48 Benabou, S. *et al.* Solution equilibria of cytosine- and guanine-rich sequences near the promoter region of the n-myc gene that contain stable hairpins within lateral loops. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1840**, 41-52 (2014).

49 Kang, C. *et al.* Stable loop in the crystal structure of the intercalated four-stranded cytosine-rich metazoan telomere. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 3874-3878 (1995).

50 Fernández, S., Eritja, R., Aviñó, A., Jaumot, J. & Gargallo, R. Influence of pH, temperature and the cationic porphyrin TMPyP4 on the stability of the i-motif formed by the 5′-(C3TA2)4-3′ sequence of the human telomere. *Int. J. Biol. Macromol.* **49**, 729-736 (2011).

51 Escaja, N. *et al.* A minimal i-motif stabilized by minor groove G:T:G:T tetrads. *Nucleic Acids Res.* **40**, 11737-11747 (2012).

52 Nonin, S., Phan, A. T. & Leroy, J. L. Solution structure and base pair opening kinetics of the i-motif dimer of d(5mCCTTTACC): a noncanonical structure with possible roles in chromosome stability. *Structure* **5**, 1231-1246 (1997).

53 Zhang, T. *et al.* The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4875-4883 (2013).

54 Blower, M. D., Sullivan, B. A. & Karpen, G. H. Conserved organization of centromeric chromatin in flies and humans. *Dev. Cell* **2**, 319-330 (2002).

55 Widlund, H. R. *et al.* Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.* **267**, 807-817 (1997).

56 Tanaka, Y. *et al.* Human centromere protein B induces translational positioning of nucleosomes on α-satellite sequences. *J. Biol. Chem.* **280**, 41609-41618 (2005).

57 Harrington, J. J., Van Bokkelen, G., Mays, R. W., Gustashaw, K. & Willard, H. F. Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. *Nat. Genet.* **15**, 345-355 (1997).

58 Ikeno, M. *et al.* Construction of YAC–based mammalian artificial chromosomes. *Nat. Biotechnol.* **16**, 431-439 (1998).

59 Ohzeki, J., Nakano, M., Okada, T. & Masumoto, H. CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *J. Cell Biol.* **159**, 765-775 (2002).

60 Okada, T. *et al.* CENP-B controls centromere formation depending on the chromatin context. *Cell* **131**, 1287-1300 (2007).

61 Fachinetti, D. *et al.* A two-step mechanism for epigenetic specification of centromere identity and function. *Nat. Cell Biol.* **15**, 1056-1066 (2013).

62 Hudson, D. F. *et al.* Centromere protein B null mice are mitotically and meiotically normal but have lower body and testis weights. *J. Cell Biol.* **141**, 309-319 (1998).

63 Kapoor, M. *et al.* The cenpB gene is not essential in mice. *Chromosoma* **107**, 570-576 (1998).

64 Perez-Castro, A. V. *et al.* Centromeric protein B null mice are viable with no apparent abnormalities. *Dev. Biol.* **201**, 135-143 (1998).

65 Li, T. & Famulok, M. i-Motif-programmed functionalization of DNA nanocircles. *J. Am. Chem. Soc.* **135**, 1593-1599 (2013).

SUPPLEMENTARY DATA FOR:

Centromeric alpha-satellite DNA adopts dimeric structures capped by AT Hoogsteen base pairs

Miguel Garavís[1,2], Núria Escaja[3], Valérie Gabelica[4,5], Alfredo Villasante[2*] and Carlos González[1*]

[1]Instituto de Química Física Rocasolano, CSIC, Serrano 119, 28006, Madrid, Spain. [2]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Nicolás Cabrera 1, 28049 Madrid, Spain. [3]Departament de Química Orgànica and IBUB, Universitat de Barcelona, Martí i Franquès 1-11, 08028 Barcelona. [4]Univ. Bordeaux, ARNA Laboratory, IECB, 2 rue Robert Escarpit F-33600 Pessac, France. [5]Inserm, ARNA Laboratory, 146 Rue Leo Saignat, F33000 Bordeaux, France.

Tables and Figures mentioned in the main text:

Supplementary Tables

Table S1: Assignment list of d(TCCTTTTCCA) (HS2) at pH 3.6, *T*=5ºC. [HS2] = 2.9 mM

	H1/H3	H42/H62	H41/H61	H6/H8	H5/Me/H2	H1'	H2'	H2"	H3'	H4'	H5'	H5"
T1	11.36	-	-	7.77	1.73	5.40	2.51	1.98	4.63	3.78	3.66	
C2	15.68	9.97	8.58	7.91	5.87	6.56	2.12	2.60	4.84	4.20	n.a.	
C3	15.32	9.52	8.36	7.90	6.00	6.43	2.61		4.89	4.42	n.a.	
T4	n.a.	-	-	7.79	1.94	6.45	2.35	2.55	4.60	4.55	4.17/3.95	
T5	n.a.	-	-	7.75	1.97	6.37	2.27	2.79	4.75	4.48	4.16/4.50	
T6	12.77	-	-	7.62	1.92	6.04	2.08	2.35	4.86	3.72	3.96/3.93	
T7	11.27	-	-	7.83	1.89	6.35	2.60	2.16	4.90	4.33	n.a.	
C8	15.68	8.90	8.15	7.74	5.91	6.44	2.00	2.54	4.71	4.28	n.a.	
C9	15.32	9.05	8.24	7.74	6.00	6.14	2.17	2.50	4.62	4.25	4.14/3.96	
A10	n.o.	8.76	n.o	8.39	8.47	6.51	2.67	2.97	4.73	4.29	4.02	

**Chemical shifts of imino protons of T4 and T5 are: 11.20 and 11,10 ppm, but it was not possible to assign them.

n.o: not observed

n.a: not assigned

Table S2: Assignment list of d(TCCTTTTCCACC) (HS1) at pH 3.6, *T*=5ºC. [HS1] = 1.5 mM

	H1/H3	H42/H62	H41/H61	H6/H8	H5/Me/H2	H1'	H2'	H2"	H3'	H4'	H5'	H5"
T1	11.32	-	-	7.73	1.71	5.39	1.96	2.48	4.61	3.67	3.76/3.57	
C2	15.67	10.03	8.60	7.92	5.87	6.57	2.11	2.62	4.85	4.43	4.07	
C3	15.30	9.49	8.37	7.91	5.99	6.43	2.64		4.89	n.a.	n.a.	
T4	n.a.	-	-	7.77	1.93	6.47	2.37	2.57	4.61	4.53	4.20/3.98	
T5	n.a.	-	-	7.74	1.96	6.39	2.28	2.80	4.76	4.51	n.a.	
T6	12.60	-	-	7.65	1.95	6.00	2.09	2.39	4.88	3.73	3.99/4.04	
T7	11.20	-	-	7.81	1.85	6.36	2.14	2.60	4.89	4.34	4.20	
C8	15.67	8.84	8.14	7.73	5.91	6.43	2.01	2.57	4.71	4.30	n.a.	
C9	15.30	9.06	8.26	7.75	6.01	6.13	2.17	2.57	4.63	4.27	n.a.	
A10	n.o.	8.93	8.75	8.47	8.48	6.42	2.92	3.05	4.77	4.52	4.42	
C11	n.o.	9.52	8.38	8.23	6.29	6.14	2.39	2.65	4.77	4.42	4.10	
C12	n.o.	9.37	8.49	7.81	6.10	5.98	2.15	2.41	4.41	4.89	n.a.	

**Chemical shifts of imino protons of T4 and T5 are: 11.25 and 11.12 ppm, but it was not possible to assign them.

n.o: not observed

n.a: not assigned

Table S3: Assignment list of d(TTCCTTTTCTACCATAG) (HST) at pH 4.5, T=5ºC
Buffer conditions: 25 mM phosphate buffer, 100 mM NaCl

	H1/H3	H42/H62	H41/H61	H6/H8/H2	H5/Me	H1'	H2'	H2"	H3'	H4'	H5'	H5"
T0		-	-	7.64	1.85	6.19	2.22	2.51	4.75	n.a	n.a	
T1		-	-	7.71	1.87	6.31	2.48		4.91	n.a	n.a	
C2	15.18	8.79	7.51	7.71	5.68	6.35	1.58	2.25	n.a.	n.a	n.a	
C3	15.41	8.71	7.95	7.44	5.69	6.11	1.94	2.27	4.48	n.a	n.a	
T4	13.08	-	-	7.28	1.70	6.15	2.19	2.69	4.72	n.a	n.a	
T5		-	-	7.79	1.98	6.40	2.31	2.68	4.94	n.a	n.a	
T6		-	-	7.70	1.92	6.17	2.40		4.94	n.a	n.a	
T7		-	-	7.48	1.56	5.83	1.95	2.28	4.76	n.a	n.a	
C8	n.o.	9.23	8.71	7.95	6.12	6.10	2.34	2.62	4.80	n.a	n.a	
T9		-	-	7.57	1.71	6.17	2.41	2.80	4.77	n.a	n.a	
A10	-			8.17	-	6.62	2.28	2.61	5.18	n.a	n.a	
C11	15.18	8.84	7.85	7.75	6.08	6.28	2.05	2.40	4.84	n.a	n.a	
C12	15.41	9.62	8.29	7.62	5.84	n.a	n.a.	n.a.	4.81	n.a	n.a	
A13	-	6.34	6.90	8.29	-	6.28	2.45	2.88	4.91	n.a	n.a	
T14		-	-	7.39	1.59	5.81	2.36	2.40	4.86	n.a	n.a	
A15	-	n.o	n.o	7.87	-	6.14		2.68	4.79	n.a	n.a	
G16	n.o	n.o	n.o	7.11	-	5.93	2.41	2.59	4.68	n.a	n.a	

**Chemical shifts of imino protons at 10.55, 11.22 and 11.43 ppm correspond to thymines, but it was not possible to assign them.

n.o: not observed

n.a: not assigned

Table S4: Experimental constraints and calculation statistics of d(TCCTTTTCCA) (HS2)

* All except thymines 4 and 5

Experimental distance constraints		
Total number	300	
intra-residue	96	
sequential	46	
range > 1	158	
Intra-subunit	186	
Inter-subunit	114	
RMSD (Å)		
all well-defined* bases	0.3±0.1 Å	
all well-defined* heavy atoms	0.8±0.1 Å	
backbone	0.9±0.1 Å	
all heavy atoms	1.1±0.3 Å	
Residual violations	Average	Range
Sum of violation (Å)	12.5	12.0 - 13.3
Max. violation (Å)	0.71	0.50 - 0.90
NOE energy (kcal/mol)	65	55 - 75
Total energy (kcal/mol)	-1230	-1442 - -1238

Table S5. Average dihedral angles and order parameters of the dimeric structure of **HS2**

Residue	Pseudorotation		α		β		Υ		δ		ε		ζ		χ	
	Phase	Ampli.	Average	Order param.	Average	Order param.	Average	Order param.	Average	Order param.	Average	Order param.	Average	Order param.	Average	Order param.
T1	30	33	-	-	-177	0.3	80	0.6	100	1	-68	0.9	-	-	-151	1
C2	50	37	-30	0.3	-176	1	118	0.6	92	1	-174	0.3	-160	1	-155	1
C3	7	34	-36	0.4	175	1	107	0.7	101	1	-109	1	-144	0.3	-117	1
T4	154	35	-114	0.9	175	1	-173	1	160	1	-163	0.9	-69	0.8	-167	1
T5	113	38	-166	0.8	163	1	-106	1	105	1	-124	0.9	-106	0.8	-120	0.6
T6	126	35	-142	0.5	176	0.9	53	0.9	155	1	-65	1	-160	0.9	-161	1
T7	141	36	-103	0.6	167	1	88	0.7	130	1	-89	1	-102	1	-113	1
C8	109	35	-47	0.2	171	1	100	0.5	125	1	-136	0.7	-153	1	-92	1
C9	46	40	111	0.7	-170	1	-152	1	105	1	142	0.4	-175	0.7	-84	1
A10	134	44	-26	0.9	172	0.9	54	0.8	122	1	-	-	-38	0.5	-107	1
T1'	28	32	-	-	167	1	98	0.4	98	1	-61	0.7	-	-	-152	1
C2'	47	36	-59	0.8	-	-	89	1	95	1	-150	0.8	-162	1	-152	1
C3'	6	37	-86	0.7	-175	0.5	86	0.8	101	1	-104	1	-100	0.8	-117	1
T4'	154	36	-110	0.8	-177	1	-175	1	162	1	-165	1	-78	0.9	-167	1
T5'	99	34	-162	0.8	175	1	-112	1	95	1	-108	1	80	0.9	-91	0.7
T6'	163	38	-172	0.3	172	1	55	0.9	151	1	-67	1	-171	0.9	-165	1
T7'	140	40	67	0.2	171	1	151	0.5	133	1	-79	1	96	1	-115	1
C8'	143	34	62	0.5	177	0.9	-178	0.7	137	1	-112	0.7	-52	0.9	-96	1
C9'	50	36	98	0.7	166	1	-156	1	107	1	119	0.3	-178	0.8	-86	1
A10'	134	43	-59	0.5	173	1	37	0.9	122	1	-	-	-13	0.3	-105	1

Supplementary Figures

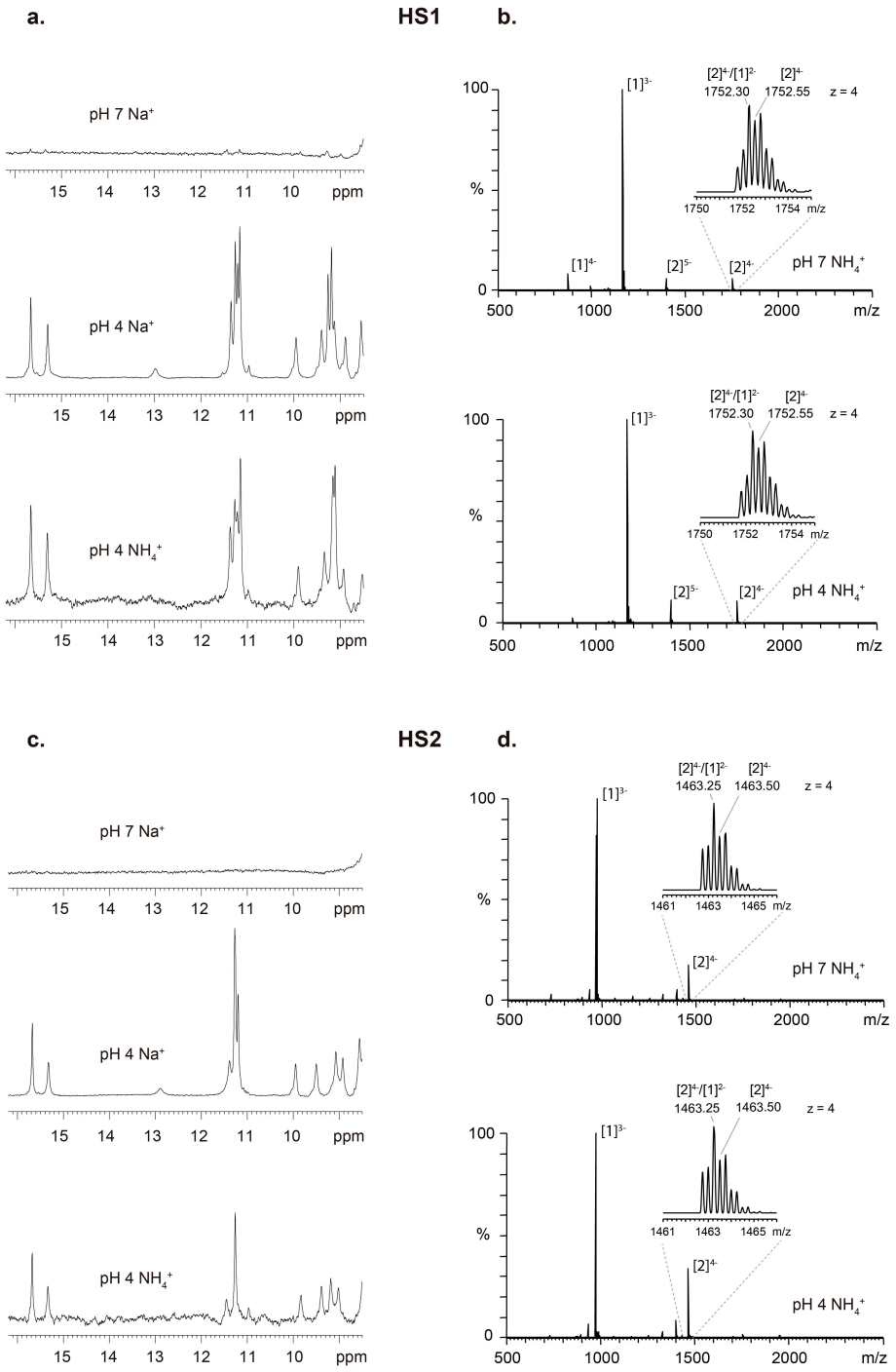


***Figure S1**. NMR spectra (a and c) and MS spectra (b and d) of HS1 and HS2 under NH_4^+ (100 mM NH_4OAc, [Oligo] = 100 µM]) and Na^+ (25 mM NaPi, 100 mM NaCl, [Oligo] = 2 mM) conditions. Zoom views of the dimer peaks showing the isotopic distribution are displayed in the insets of the panels b and d (separation between two consecutive ^{13}C isotopes in the isotopic distribution corresponds to m/z equals to 1/z, consequently in main isotopic distribution the z value is 4, and the mass is that of a dimer. The uneven isotopic profile indicates that doubly charged monomer also contributes to the signal).*

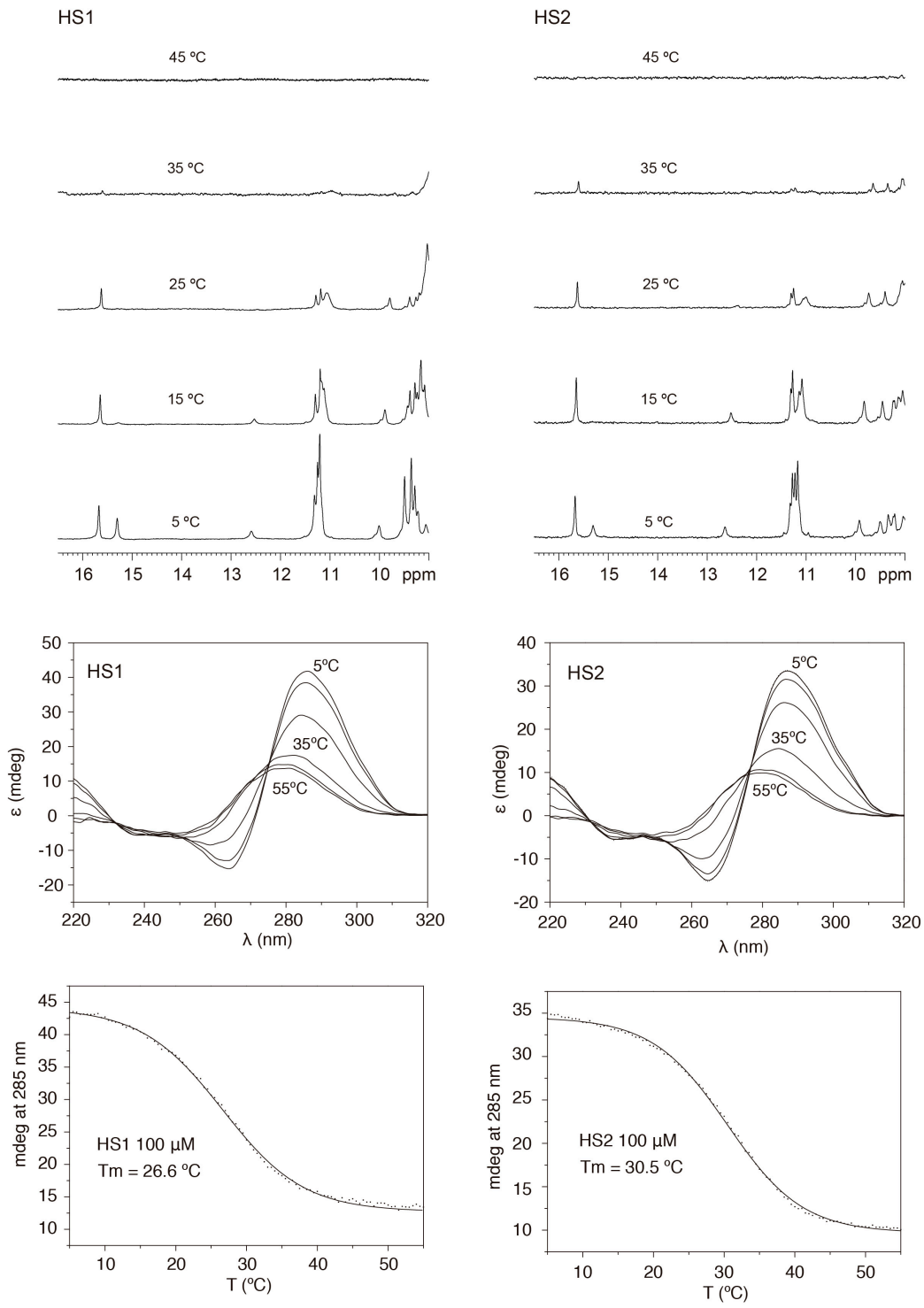


***Figure S2**. NMR and CD melting experiments of HS1 (left) and HS2 (right). NMR spectra were recorded in H_2O at 1.5 mM oligonucleotide concentration. CD experiments were performed in 25 mM sodium phosphate buffer, pH 4.0,100 mM NaCl at 100 µM oligonucleotide concentration. Tm values of HS1 and HS2 are shown in the inset of bottom panels.*

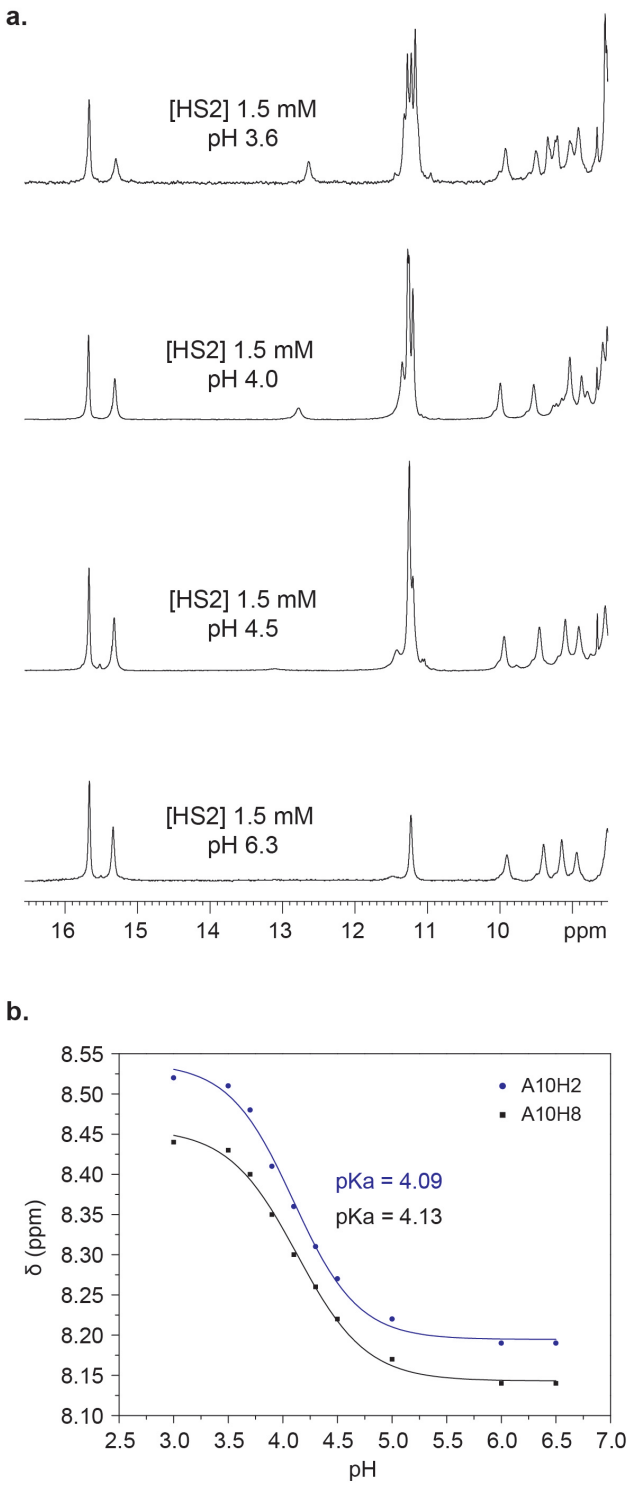


***Figure S3**. (a). [1]H NMR spectra in 90/10 H_2O/D_2O showing the imino region of HS2 at different pHs, 1.5 mM oligonucleotide concentration and T = 5ºC. (b). Titration curves of A10H2 (blue) and A10H8 (black) showing the variation of chemical shift (δ) at different pH values. The midpoints shown in the inset were calculated by fitting the data to a sigmoidal curve.*

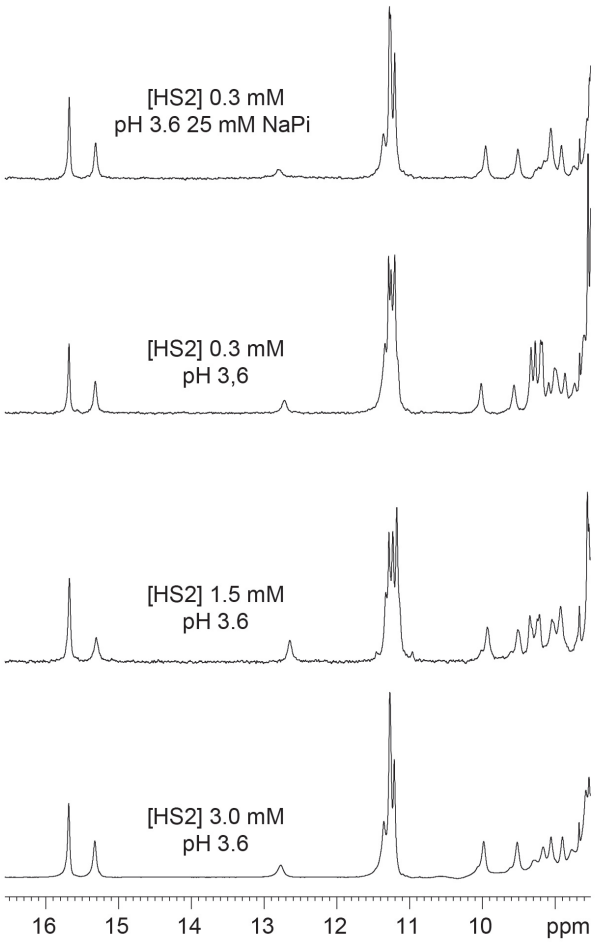


Figure S4. *^{1}H NMR spectra in 90/10 H_2O/D_2O showing the imino region of HS2 at different oligonucleotide concentrations, pH 3.6 and T = 5 ºC.*

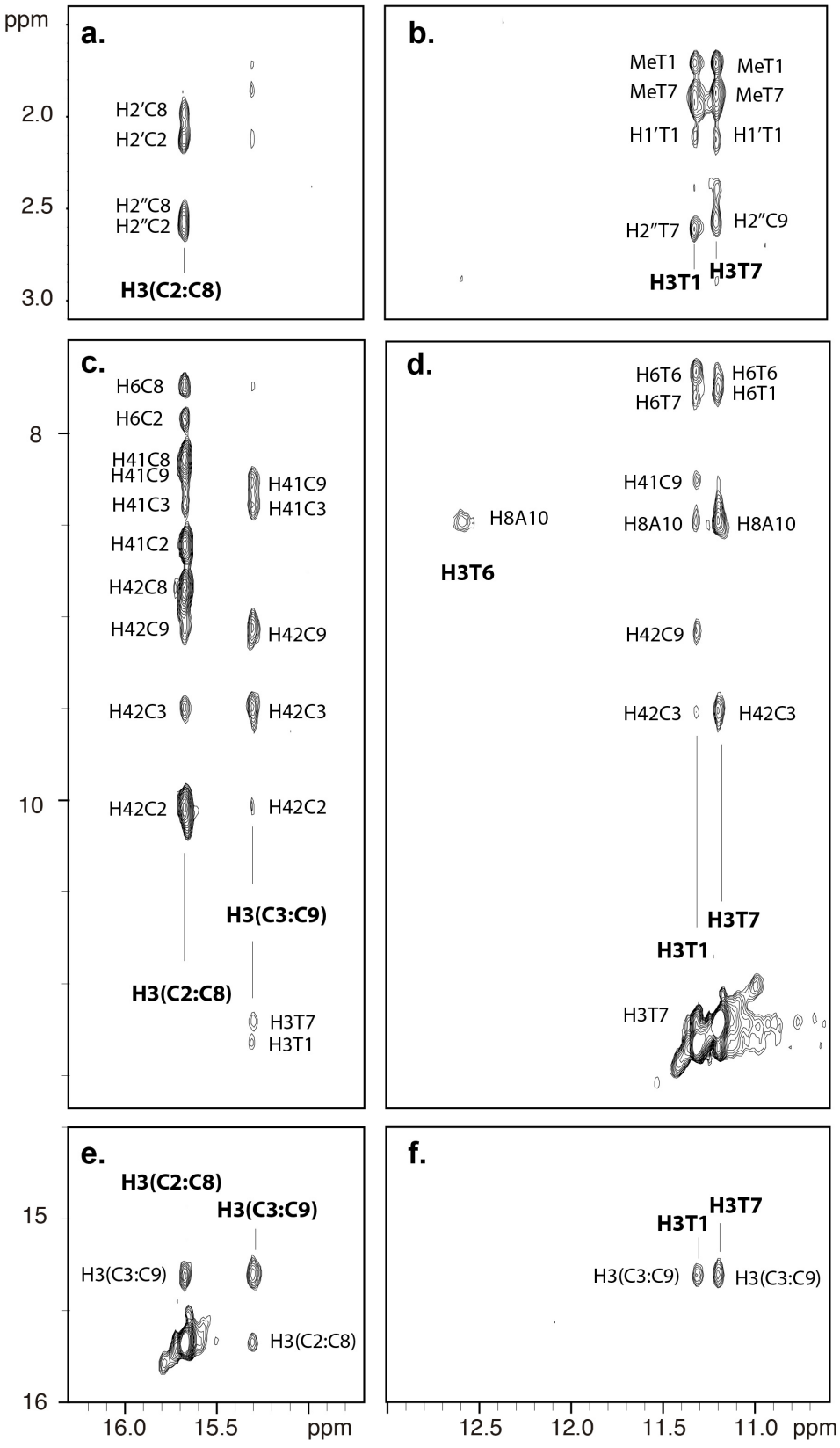


***Figure S5**. NOESY spectrum of HS1 at pH 3.6 and T=5 ºC in 90/10 H_2O/D_2O ([HS1]=1.5 mM) showing cross peaks involving the imino protons of hemiprotonated C:C$^+$ base pairs (a, c and e) and the imino protons of T1, T6 and T7, involved in T:T and Hoogsteen A:T base pairs (b, d and f).*

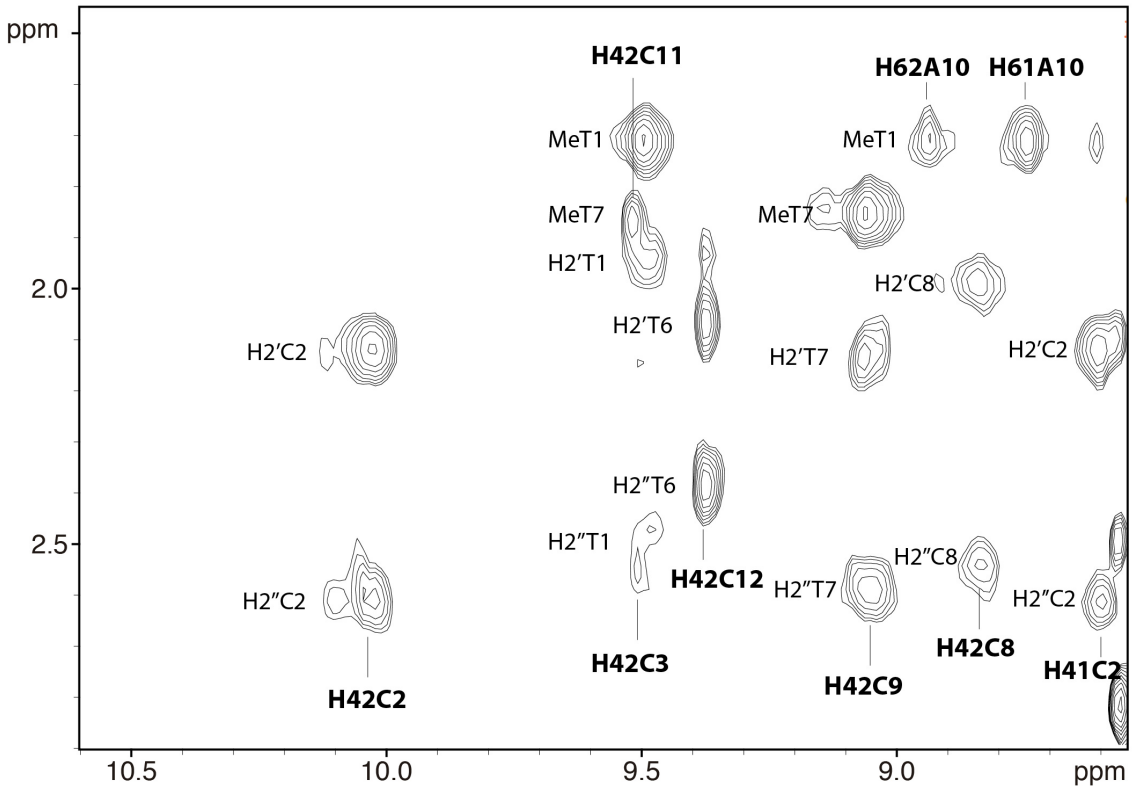


Figure S6. *NOESY spectrum of HS1 at pH 3.6 and T=5 ºC in H_2O ([HS1]=1.5 mM) showing cross peaks between H2'/H2'' and methyl protons with some of the amino protons of cytosines. These cross-peaks are characteristic of i-motif structures and occur between residues facing their 3' edges across the wide groove. The spectrum also shows cross peaks involving methyl of T1 and the amino protons of adenine.*

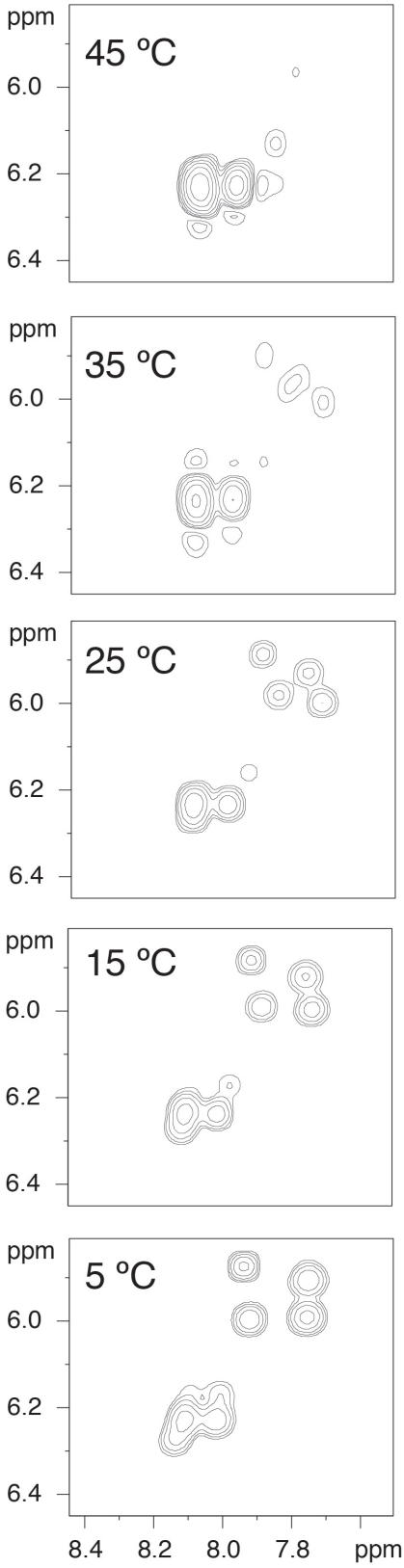


Figure S7. *TOCSY spectra of HS2 showing the cross peaks H5-H6 of cytosines. Signals between 8.0 and 8.2 ppm correspond to the unfolded species and their intensities increase at higher temperature. The opposite effect is observed for the signals of the folded species, which progressively disappear with the increase of temperature, being negligible over the melting temperature.*

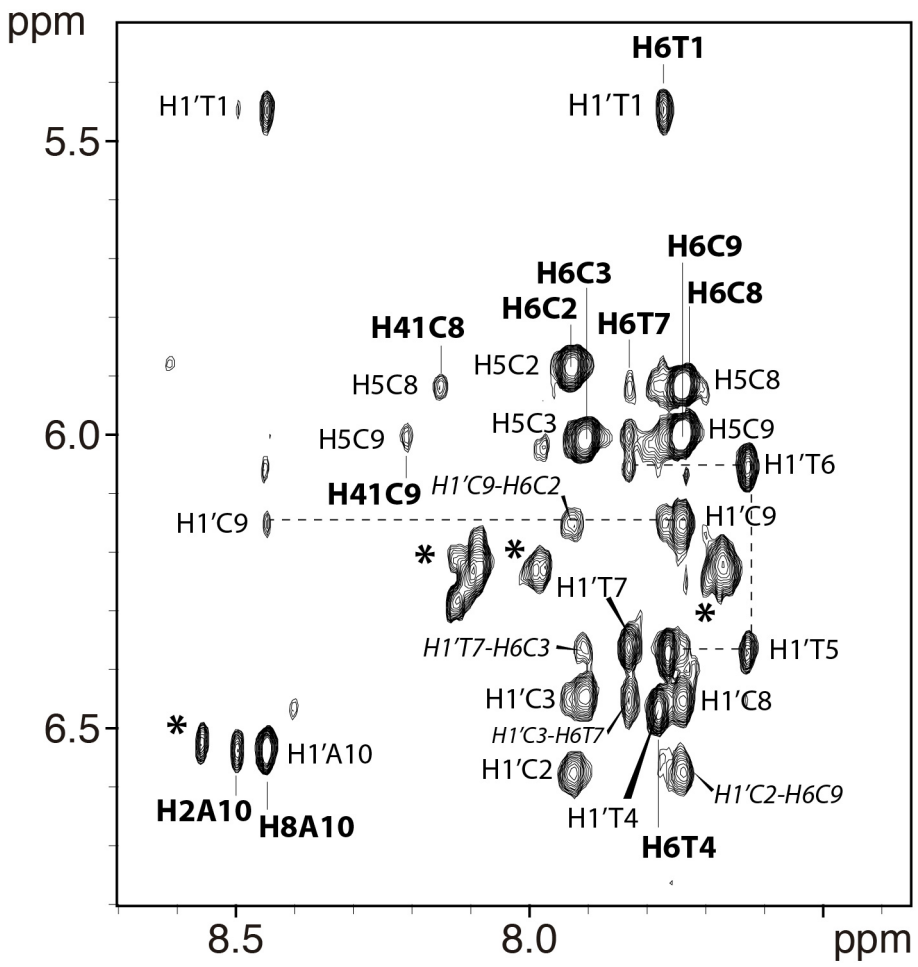


***Figure S8**. NOESY spectrum of **HS2** in D_2O, showing the H1'-aromatic region. The connectivity between T5, T6 and T7 loop residues and between C9 and A10 is indicated. Cross-peaks involving exchangeable protons C8H41 and C9H41 are observable due to residual H_2O in the sample. Signals from the unfolded species (see Figure S7 legend) are marked with (*).*

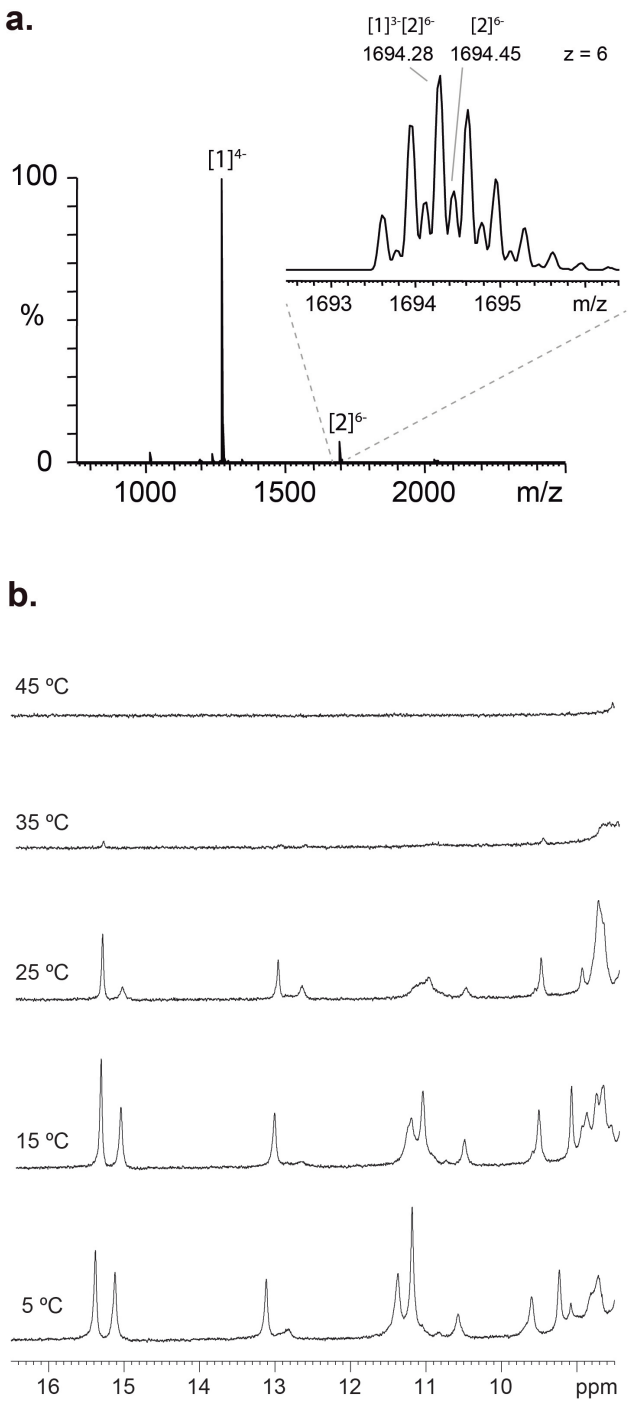


***Figure S9**. a) Mass spectrum of **HST** at 100 µM (100 mM NH4OAc buffer, pH 7). The isotropic distribution of the peak marked with the asterisk is showed in the inset of the MS spectrum. b) 1D NMR spectra showing the imino region of HST at pH 3.6 in 90/10 H_2O/D_2O at different temperatures ([HST] = 2 mM).*

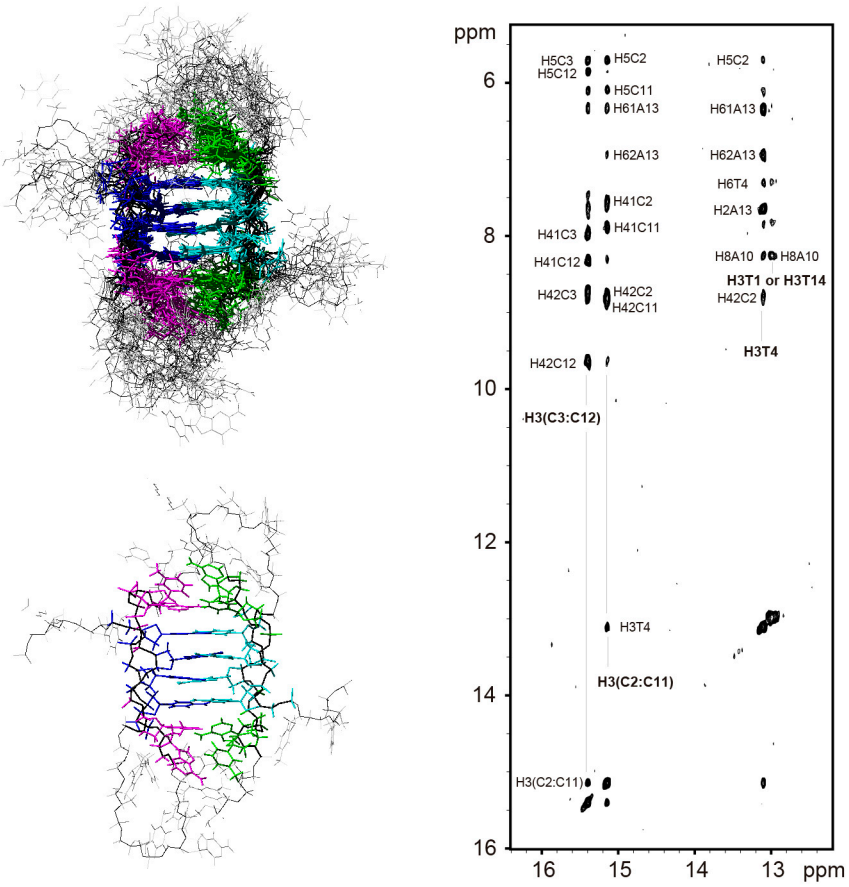


Figure S10. *(Right) Imino region of the NOESY spectra of* ***HST*** *at 2 mM (150 ms mixing time). Buffer conditions: 25 mM phosphate buffer, 100mM NaCl, pH 4, T = 5 ºC. Left) NMR-based model of the structure of* ***HST.*** *Ensemble of 10 structures (top), and average structure (bottom). Colour code: Cytosines in the two sub-units are shown in blue and cyan, respectively; well-defined thymines and adenines in the two sub-units are shown in magenta and green, respectively; no well-defined residues in grey; and backbone in black.*

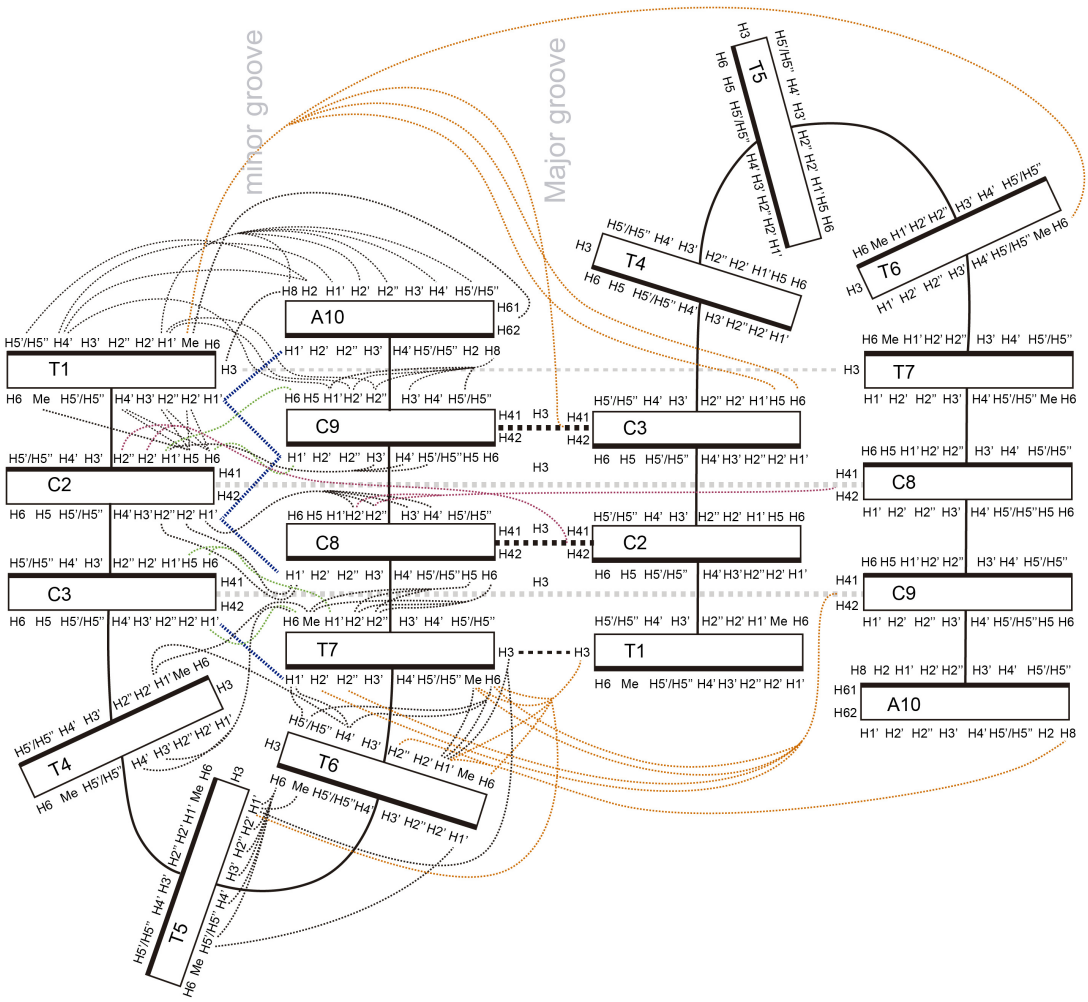


Figure S11. *Proton connectivity map of HS2 d(TCCTTTTCCA). Thick black lines mark the 5' faces of the bases. Thick black and grey dashed lines represent connectivities from C:C$^+$ base pairs. Thin black and orange dashed lines correspond to intra- and inter-strand connectivities, respectively. Thick blue dashed lines correspond to the characteristic H1'-H1' cross peaks between stacked bases through the minor groove. Thin red dashed lines represent the characteristic H2'/H2''-amino cross-peaks between cytosines facing their 3' faces through the major groove. Thin green dashed lines represent the reciprocal H1'-H6 connectivities between C3-T7 and C2-C9 residues.*

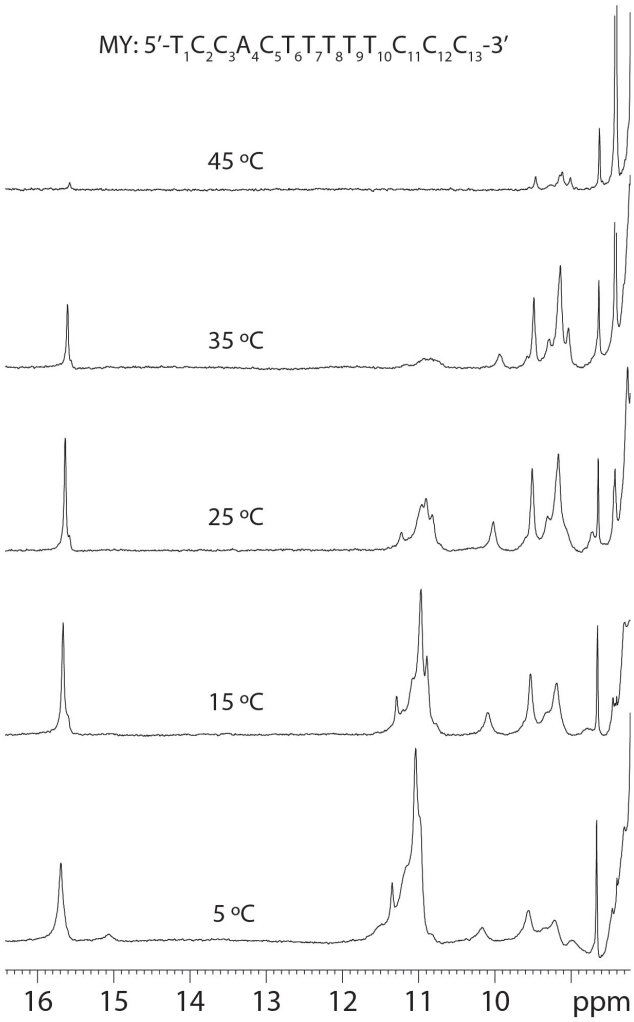


Figure S12. *1D NMR spectra showing the imino region of the MY sequence at pH 4.0 in 90/10 H_2O/D_2O at different temperatures. The MY sequence belongs to the mouse Y centromeric satellite. This Y-specific satellite DNA is a highly diverged minor satellite-like sequence. The MY sequence appears in a position equivalent to the one of the CENP-B box in the centromeric minor satellite repeat.*

Artículo 5.

The structure of an endogenous *Drosophila* centromere reveals the prevalence of tandemly repeated sequences able to form i-motif

Miguel Garavís, María Méndez-Lago, Valerie Gabelica, Siobhan L. Whitehead, Carlos González and Alfredo Villasante

La estructura del centrómero endógeno de *Drosophila* revela el predominio de secuencias repetidas en tándem capaces de formar *i-motif*

Miguel Garavís, María Méndez-Lago, Valerie Gabelica, Siobhan L. Whitehead, Carlos González and Alfredo Villasante

Los centrómeros de la mayoría de los organismos eucariotas están compuestos por secuencias que se repiten un número elevado de veces (DNA satélite) y por elementos transponibles que, con una menor frecuencia, también aparecen repetidos. Es, precisamente, la naturaleza repetitiva de estas secuencias lo que dificulta en gran medida la determinación de la secuencia completa de los centrómeros.

Existen evidencias que sugieren que cada centrómero endógeno de *Drosophila melanogaster* está compuesto por diferentes DNA satélites. La región centromérica del cromosoma 3 contiene el DNA satélite dodeca, una secuencia de 11/12 pares de bases que tiene una distribución asimétrica de residuos de guanina y citosina entre las dos hebras. En este trabajo se presenta la estructura molecular del centrómero del cromosoma 3. Se trata de la primera estructura molecular de un centrómero endógeno de *Drosophila melanogaster*. El DNA satélite dodeca se encuentra repartido en dos bloques mayoritarios siendo diferente la distribución de las secuencias de 11 y 12 pares de bases en cada bloque. Además, el mapa físico muestra la presencia de elementos transponibles y de dos duplicaciones segmentales. Por otro lado, se pudo determinar la localización del DNA satélite 10 bp resultando que éste se encuentra físicamente unido al DNA satélite dodeca y, por lo tanto, en una posición diferente a la que se le atribuía en estudios anteriores.

La paradoja del centrómero surge del hecho llamativo de que las proteínas centroméricas estén conservadas a pesar de que las secuencias de DNA del centrómero evolucionan rápidamente a través de procesos inevitables de recombinación que tienen lugar entre secuencias repetidas. La existencia de un motivo estructural de DNA independiente de secuencia, podría ser una explicación a esta paradoja. En este trabajo se ha llevado a cabo el estudio estructural de secuencias derivadas del DNA satélite dodeca con el fin de explorar la capacidad de estas secuencias de plegarse formando estructuras no canónicas de DNA. El análisis mediante RMN, CD y Espectrometría de masas de secuencias de la hebra rica en guaninas y de la hebra rica en citosinas muestra que las primeras se pliegan formando horquillas de DNA y que las últimas forman estructuras diméricas tipo i-motif a pH ácido. Por último se analizó una secuencia rica en citosinas presente en el DNA satélite centromérico 359 bp del cromosoma X observándose que dicha secuencia también es capaz de plegarse formando estructuras tipo i-motif a pH ácido. Estos datos, junto con los resultados obtenidos en el estudio de las secuencias centroméricas humanas nos invitan a plantear un posible papel de la estructura i-motif en la formación de la heterocromatina centromérica.

Aportación personal al trabajo: Preparación de las muestras de DNA para su posterior estudio por RMN, espectrometría de masas y dicroísmo circular. Adquisición de los experimentos de RMN, espectrometría de masas y dicroísmo circular. Escritura y discusión del manuscrito.

The structure of an endogenous *Drosophila* centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs.

Miguel Garavís[1,2], María Méndez-Lago[1,6], Valérie Gabelica[3,4], Siobhan L. Whitehead[5], Carlos González[2] * and Alfredo Villasante[1] *

[1]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Nicolás Cabrera 1, 28049 Madrid, Spain. [2]Instituto de Química Física Rocasolano, CSIC, Serrano 119, 28006 Madrid, Spain. [3]Univ. Bordeaux, ARNA Laboratory, IECB, 2 rue Robert Escarpit F-33600 Pessac, France. [4] Inserm ARNA Laboratory, 146 rue Leo Saignat, F-33000 Bordeaux, France. [5]The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom.

[6]Present address: Centro Nacional de Análisis Genómico, Baldiri Reixac 4, 08028 Barcelona, Spain.

Abstract

Centromeres are the chromosomal loci at which spindle microtubules attach to mediate chromosome segregation during mitosis and meiosis. In most eukaryotes, centromeres are made up of highly repetitive DNA sequences (satellite DNA) interspersed with middle repetitive DNA sequences (transposable elements). Despite the efforts to establish complete genomic sequences of eukaryotic organisms, the so-called 'finished' genomes are not actually complete because the centromeres have not been assembled due to the intrinsic difficulties in constructing both physical maps and complete sequence assemblies of long stretches of tandemly repetitive DNA. Here we show the first molecular structure of an endogenous Drosophila centromere and the ability of the C-rich dodeca satellite strand to form dimeric i-motifs. The finding of i-motif structures in simple and complex centromeric satellite DNAs leads us to suggest that these centromeric sequences may have been selected not by their primary sequence but by their ability to form noncanonical secondary structures.

Introduction

Centromere sequences evolve rapidly due to inevitable recombination processes undergone by tandemly repeated sequences but many centromere proteins are conserved[1]. This paradox could be explained by the presence of a conserved sequence-independent structural motif, rather than a particular sequence motif[2]. The maintenance of a centromere-specific motif through evolution could be driven by molecular co-evolution of centromeric DNA and centromeric proteins. Importantly, in support of this evolutionary hypothesis, it has been shown that the centromere-specific histone H3 variant CENP-A (CID in *Drosophila*) has evolved in concert with centromeric satellite DNAs[3,4].

In addition, recent study has reported that the vertebrate CENP-A-specific histone chaperone HJURP (Holliday junction recognition protein) has a DNA binding domain required for CENP-A deposition at centromeres[5]. Since HJURP preferentially binds to noncanonical DNA structures[6], the CENP-A loading mechanism might be mediated by a conserved structural motif.

Although a centromeric structural motif might be sufficient to direct the formation of centromeric chromatin on its own, the episodic occurrence of centromere activity associated with

noncentromeric sequences, neocentromeres[7,8], and the frequent inactivation/reactivation of centromeres[9-14] indicates that centromere specification involves both genomic competency and epigenetic mechanisms[15]. It is now recognized that CENP-A-containing nucleosomes provide the epigenetic mark to establish the centromere-specific chromatin[16-18], and that the centromeric chromatin contains blocks of CENP-A nucleosomes interspersed with blocks of canonical histone H3 nucleosomes[19]. However, the folding of this centromeric chromatin is still not clear and is of much controversy[20,21].

Although there is evidence to suggest that each *Drosophila melanogaster* endogenous centromere is made up of different simple and complex satellite DNAs[22-27], their molecular structure has yet to be determined. The centromeric region of chromosome 3 of *D. melanogaster*, as well as the centromeric region of chromosomes 2 and 3 of the sibling species *Drosophila simulans* and *Drosophila mauritiana*, contains dodeca satellite 11/12 bp tandem repeats (CCCGTACTGGT/CCCGTACTCGGT) showing asymmetric distribution of guanine and cytosine residues such that one strand is predominantly G-rich and the other C-rich[23,28,29].

In order to fully understand the structural and functional aspects of centromeres, it is important to elucidate the types of secondary DNA structures that can be formed by their constituent repeat units. Hence, we determined that the G-rich dodeca satellite strand is able to fold into very stable intramolecular hairpin structures that are stabilized by the formation of noncanonical G-A pairs[30], and recently we have shown that not only the type B monomer of the human centromeric alpha-satellite[31] but also the type A are able to form dimeric i-motif structures (Garavís et al., submitted). The i-motif is a four-stranded intercalated structure formed by the association of two parallel duplexes combined in an antiparallel fashion by forming intercalated hemi-protonated $C:C^+$ base pairs[32,33]. As i-motif formation requires protonation of cytosines [34], these structures are more stable at acidic pH, although, depending on particular C-rich sequences, they can fold close to neutral pH[35]. I-motifs can also exist at neutral pH under molecular crowded conditions[36] and under transcriptionally induced negative superhelicity[37].

Results and Discussion

Since centromere specification may rely on centromeric structural motifs under the control of epigenetic mechanisms, knowledge of the fine structure of the endogenous *D. melanogaste*r centromeres is required to elucidate the potential formation of noncanonical centromeric DNA structures. However, due to the repetitive nature of centric heterochromatin, the endogenous centromeres remain poorly represented in the new *D. melanogaster* Release 6 reference genome sequence (Hoskins et al., submitted).

The heterochromatin of *D. melanogaster* has been subdivided into 61 distinct cytological regions[38], and the primary constriction of the third chromosome localizes asymmetrically within the h53 region[38]. Previous work has shown that the dodeca satellite DNA hybridized very close to the

primary constriction of the third chromosome, but extending to the right arm[29], and that, occasionally, two hybridization signals can be seen in prometaphase chromosomes[28]. Moreover, the cytological analysis of the free chromosome arm *F(3R)1* has shown that the amount of dodeca satellite on the right arm can be reduced without compromising chromosome segregation[29]. Initial studies of the long-range structure of the dodeca satellite DNA were reported by Losada et al. (2000). Common-cutting restriction enzymes that do not cut within dodeca satellite DNA revealed that most of the dodeca satellite was organized in two major blocks. In addition, the distribution of restriction sites in the long-range map suggested that the region between the blocks consists of complex DNA sequences, while the flanking region of one of the blocks is likely to contain another putative block of repetitive DNA yet undescribed.

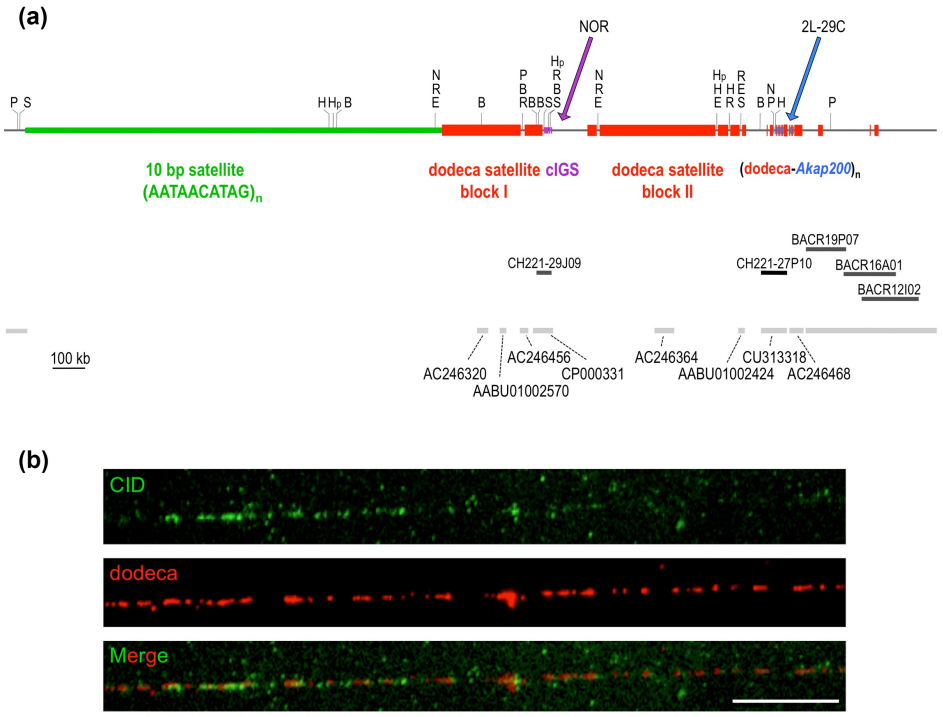


Figure 1. *Structure of the centromere of the third chromosome of D. melanogaster. (a) 2.5 Mb physical map across the centromere of chromosome 3. The regions containing the 10 bp satellite repeats and the dodeca satellite repeats appear in green and red, respectively. The segmental duplications from the NOR and from region 2L-29C are indicated in purple and blue, respectively. The position and GenBank number of the eight centromeric scaffolds are also indicated. Abbreviations are: B, BamHI; H, BssHII; E, BstEII, R, EcoRI; N, NaeI; Hp, HpaI; P, PmeI, S, SwaI. (b) Extended chromatin fibers from S2 cells were processed for immunofluorescence with an anti-CID antibody followed by FISH with the dodeca satellite oligo probe. A representative image showing CID immunostaining (green) overlapping with dodeca (red) is shown. Of a total of 43 chromatin fibers stained with the anti-CID antibody, 14 showed co-localization with dodeca, a proportion in agreement with the karyotype of the polyploid S2 cells. CID signals do not encompass all dodeca satellite repeats. Scale bar is 5 µm.*

Since a complete physical map across the centromere should extend from chromosome arm 3L to chromosome arm 3R, we set out to isolated bacterial artificial chromosome (BAC) clones that contain dodeca satellite, and to construct a comprehensive map around the dodeca satellite blocks

using 20 restriction enzymes that do not occur within the dodeca satellite. Single and double genomic digests were size-fractionated by pulsed-field gel electrophoresis (PFGE) using a "Waltzer" apparatus[39], which gives sharp resolution up to 2 Mb (representative digests are shown in Supplementary Fig. S1).

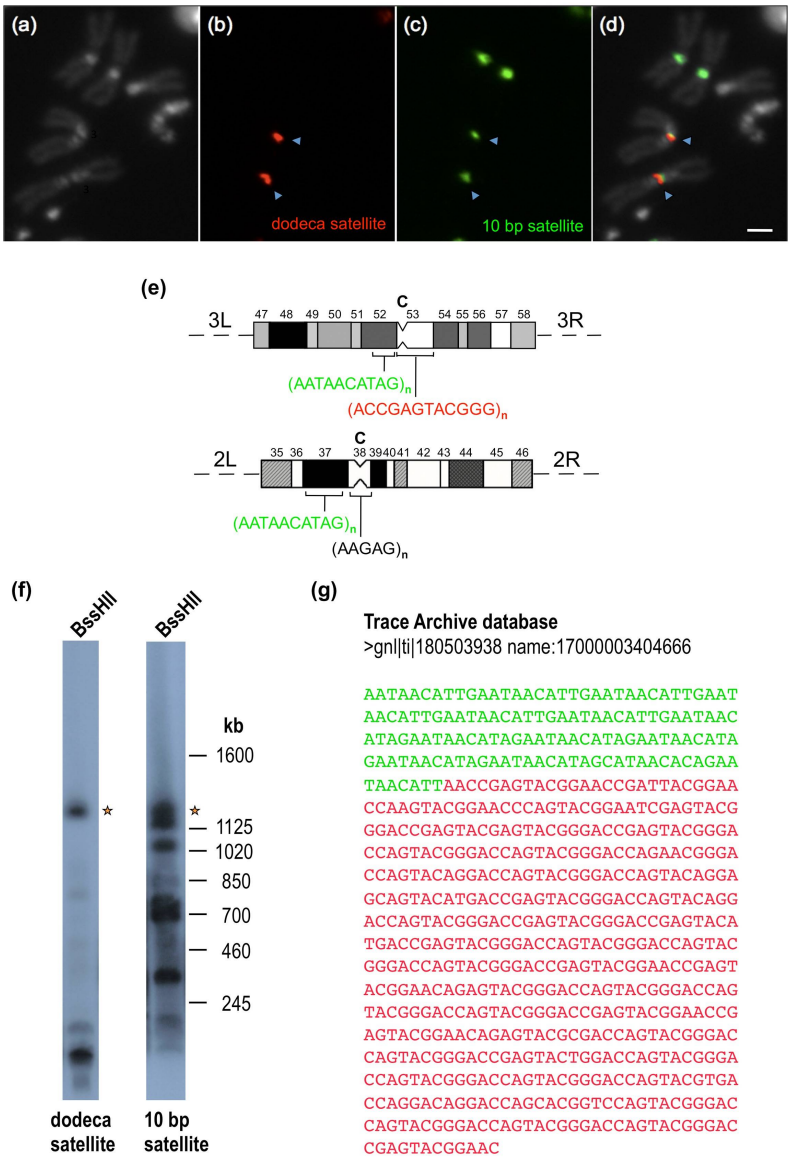


Figure 2. *The 10 bp satellite DNA localizes on the third chromosome at h52p instead of h48. (a) Metaphase chromosomes counterstained with DAPI. (b) Hybridization signals from a dodeca satellite probe (in red). (c) Hybridization signals from a 10 bp satellite probe (in green). (d) Hybridization signals superimposed with DAPI-stained chromosomes. The Scale bar is 2 µm. (e) Diagram representing the heterochromatic regions (Gatti and Pimpinelli 1992) of chromosomes 2 (regions 35-46) and 3 (regions 47-58) showing the localization of the 10 bp (in green) and dodeca (in red) satellites. The position of the centromeres (C) is indicated. (f) High molecular weight DNA from red e embryos was digested with BssHII, electrophoresed through a 1% (w/v) agarose gel in a "Waltzer" apparatus at 150 V for 24 h with a 130 s pulse time, blotted onto a nylon filter and hybridized successively with the dodeca satellite probe pBK6E218 at 68°C and with the 10 bp satellite probe 5'-AATAACATAGAATAACATAGAATAACATAGAATAACATAGAATAACATAG-3' at 50°C. The asterisks indicate a 1.2 Mb fragment that hybridizes with both probes. (g) DNA sequence showing a junction between 10 bp satellite repeats (in green) and dodeca satellite repeats (in red).*

In order to obtain dodeca satellite clones, three *D. melanogaster* BAC libraries were screened: the RPCI-98 library generated by cloning EcoRI-digested genomic DNA and the CHORI-221 and CHORI-223 libraries generated from sheared genomic DNA. BAC end sequencing and fingerprinting data of the stronger clones were used to construct contigs, and five clones were chosen for complete sequencing: CH221-29J09 (containing rDNA intergenic spacer (IGS) related sequences, cIGS), CH211-27P10 (containing *Akap200* related sequences) and BACR19P07, BACR16A01 and BACR12I02 that were also positive for the retrotransposon *Circe*. The presence of *Circe* sequences in the centromeric region h53 had previously been shown by [40]. By combining data from the sequence of these BACs and the sequence of whole genome shotgun scaffolds containing dodeca satellite with the results of an accurate restriction site mapping of genomic DNA, it has been possible to determine the position and orientation of the first eight scaffolds in the Release 6 assembly of the chromosome arm 3R (Fig. 1a).

The dodeca satellite sequences at this endogenous centromere are organized as two adjacent major blocks, block I and block II, plus several minor blocks (Fig.1a). Interestingly, a detailed analysis of the sequence of the blocks has shown that block I has more undeca than dodeca repeats (Supplementary Fig. S2). Moreover, this centromeric region contains transposable elements and two segmental duplications: one results from a duplication of a fragment of *Akap200* (chromosome arm 2L at 29C) and subsequent amplification, and the other, located at one edge of block I, results from a duplication of IGS sequences at the nucleolus organizer region (NOR) (Fig. 1a). Nevertheless, FISH to mitotic chromosomes under low-stringency conditions with a cIGS-specific probe has not detected IGS-related sequences in the centromeric region of chromosome 2, although clear cross-hybridization signals occur at the NORs (Supplementary Fig. S3).

During our effort to identify the putative block of simple sequence DNA in the flanking region of the dodeca satellite block 1, we repeated the cytological mapping of the 10 bp satellite (AATAACATAGn) using a fluorescent probe, which improves sensitivity and resolution with respect to results obtained with tritiated probes[25]. The 10 bp satellite had been mapped by Lohe et al., (1993) to region h37 on the second chromosome (contiguous to the centromeric region h38) and to region h48 on the third chromosome (far away from the centromeric region h53). Unexpectedly, FISH experiments with dodeca and 10 bp satellite probes revealed no additional sites for the 10 bp satellite, but showed a change in its location on the third chromosome from h48 to h52p, a position which is very close to dodeca satellite (Fig. 2 a-e). To investigate further the possibility that the flanking satellite DNA corresponds to the 10 bp satellite, we asked whether the 1.2 Mb *BssHII* fragment containing both dodeca satellite and flanking satellite sequences would hybridize with the 10 bp satellite probe. To this end, genomic DNA was digested with *BssHII*, size-fractionated by PFGE, transferred to a nylon membrane, hybridized with the dodeca satellite probe and then stripped and re-hybridized with the 10 bp satellite probe. As can be seen in Figure 2 f, there is a 1.2 Mb fragment (labeled with an asterisk) that hybridizes with both probes. Finally, the junction

between 10 bp satellite and dodeca satellite sequences was found by searching the Trace Archive database (Fig. 2 g). This result indicates that the 10 bp satellite DNA is physically linked to the dodeca satellite DNA. Here, it is important to remember that PROD, a protein required for centromere condensation[41] and that specifically recognizes the 10 bp satellite[41], is located near but not in the CID-containing chromatin[42]. Therefore, the physical map constructed comprises two adjacent chromatin domains with distinct functions.

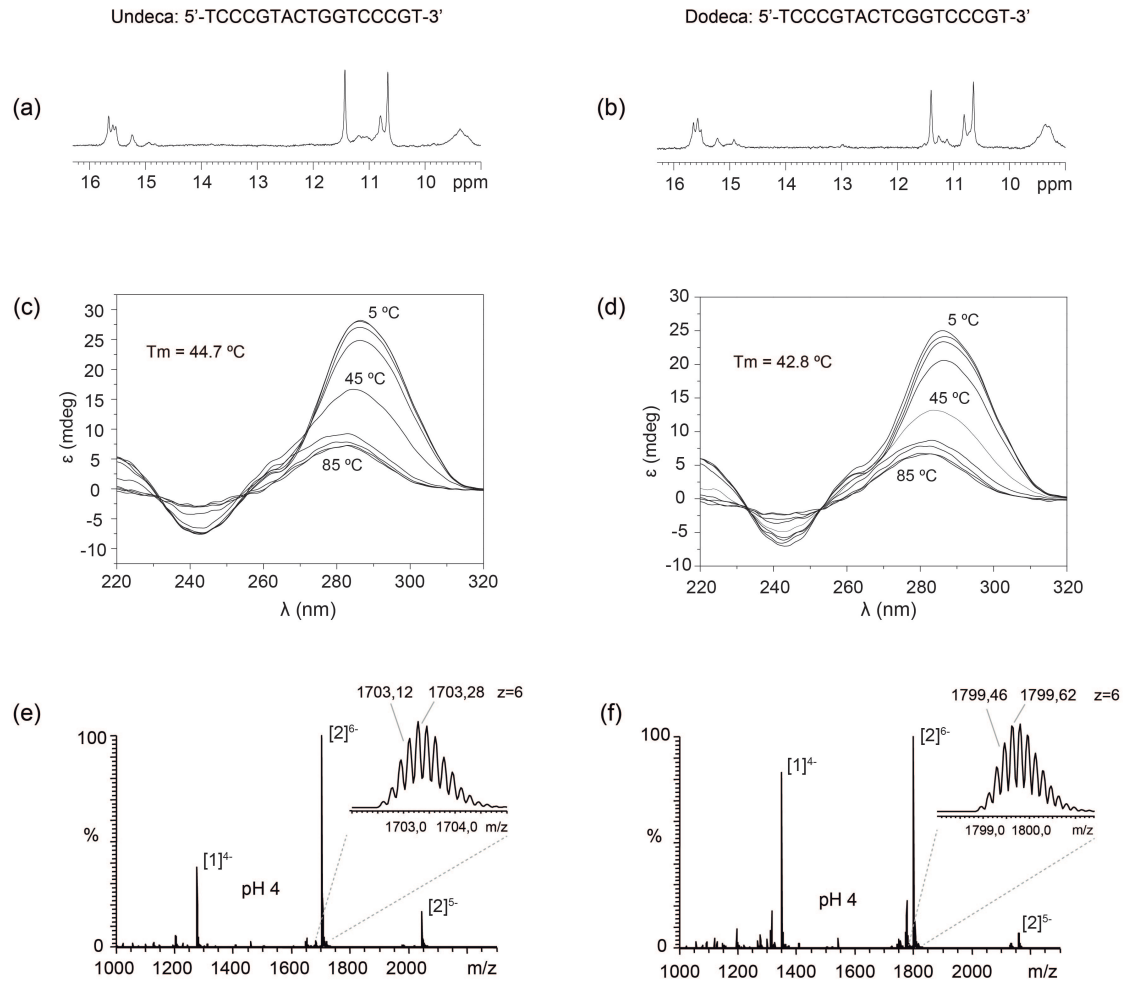


Figure 3. *The centromeric dodeca satellite DNA is able to form dimeric i-motif structures. Imino region of the NMR spectra of the C-rich strands of the undeca (a) and dodeca repeats (b). Experimental conditions: Oligo concentration = 0.8 mM, 25 mM sodium phosphate, 100 mM NaCl, T= 5ºC , pH 4. CD spectra of the C-rich strands of undeca (c) and dodeca repeats (d) at different temperatures. Oligo concentration = 100 µM, same buffer as the NMR experiments. Mass spectrometry data showing the peaks of the single stranded [1] and dimeric [2] species formed by C-rich strands of undeca (e) and dodeca (f). Buffer conditions: 100 mM NH_4OAc, spectra at pH 4. See Supplementary Fig. S6 legend for details.*

To elucidate whether CID interacts with dodeca satellite sequences, immunofluorescence-FISH experiments were performed. Thus, by using *SuUR Su(var)3-9* double mutants to suppress the normal under-replication of *Drosophila* heterochromatin during the process of polytenization, we showed that CID co-localizes on polytene chromosomes with dodeca satellite sequences[43]. To

corroborate this interaction, we increased the resolution by using extended chromatin fibers and found that anti-CID antibody and dodeca satellite signals co-localize (Fig. 1 b). Taken together, these results suggest that dodeca-satellite block I is a good candidate for the centromere region of chromosome 3.

To determine the structural behavior of the dodeca satellite DNA, several oligonucleotides containing the dodeca repeat and, its main variant, the undeca repeat were studied by NMR, circular dichroism (CD) and mass spectrometry (MS). The G-rich and the C-rich strands were analyzed under different experimental conditions (Supplementary Fig. S4 and S5). In all cases, the NMR spectra of the G-rich oligonucleotides indicate the formation of G:C base pairs, and no formation of G-tetrads is observed even at high K^+ concentrations (Supplementary Fig. S4). This is in agreement with the formation of intramolecular hairpins previously reported (Ferrer et al. 1995). However, under acidic conditions the oligonucleotides corresponding to the C-rich strand of the dodeca and undeca repeats exhibit the characteristic features of i-motif formation (sharp imino signals around 15-16 ppm) (Fig. 3 a, b and Supplementary Fig. S5). I-motif formation is confirmed by the strong positive band at 285 nm observed in the CD spectra (Fig. 3 c, d).

Mass spectrometry data clearly indicate the formation of dimeric structures (Fig. 3 e, f and Supplementary Fig. S6 a-d). CD melting experiments show that these dimeric structures are stable at pH 4.0, with melting temperatures around 42ºC for dodeca and 45ºC for undeca (Supplementary Fig. S6 e, f). These biophysical data are consistent with the self-association of these C-rich oligonucleotides through formation of dimeric i-motifs; the same secondary structures as those observed in the A and B box of the human alpha satellite (Garavís et al., submitted). All these results prompted us to check whether the 359 bp satellite DNA (centromeric DNA from the X chromosome) has the ability to form i-motifs. As can be seen in Supplementary Figure S7, the 359 bp C-rich region (found at internucleosomal linkers [44]) is also able to form i-motifs. This finding leads us to propose that these centromeric sequences may have been selected not by their primary sequence but by their ability to form noncanonical secondary structures.

Materials and Methods

Drosophila strains and cell lines

Oregon R was used as wild-type strain. The isogenic *red e* strain was used for the construction of the physical map. Standard culture conditions and media were used. *Drosophila* S2 cells were grown and maintained as described [42].

DNA analysis, sequencing and probes

High molecular weight DNAs from 0-12h *Drosophila* embryos were prepared in agarose plugs as previously described by Karpen and Spradling (1990, 1992) and Abad et al. (1992). Restriction enzyme digestions were performed following the suppliers' recommendations. DNA was analyzed

by pulsed-field gel electrophoresis using a "Waltzer" apparatus [39], and transferred to Hybond N^+ nylon filters (Amersham) in 0.4 M NaOH.

The dodeca satellite probe was pBK6E218 (Abad et al. 1992). The 10 bp satellite oligo probe was 5'-AATAACATAGAATAACATAGAATAACATAGAATAACATAGAATAACATAG-3'. The centromeric IGS (cIGS) probe (5.6 kb fragment) was obtained from BACR31J03 using the primers: cIGS-Fw: 5'-TGGCAGCGTTTTAAGGGATG-3' and cIGS-Rv: 5'-TAAGACGCCTGCAGAGAACG-3'. The PCR was carried out as described by Losada et al. (1997). The PCR product was cloned in vector pGEM-T (Promega). Plasmid probes were ^{32}P-labeled by random-priming and oligonucleotide probes were ^{32}P-labeled with T4 polynucleotide kinase. The BAC clones were sequenced at The Wellcome Trust Sanger Institute by the standard shotgun sequencing and directed finishing approach. The GenBank accession number for the sequence of BAC19P07, BAC16A01, BAC12I02, CH221-29J09 and CH221-27P10 are CU311183, CR942806, CR942807, CU463787 and CU313318, respectively.

Fluorescence in situ hybridization to mitotic chromosomes

Larval neuroblast chromosomes from Oregon R were prepared as described previously (Gatti et al. 1994). Chromosomes were counterstained with 4′,6-diamino-2-phenylindole (DAPI). The dodeca satellite oligo probe 5'-CCCGTACTGGTCCCGTACTGGTCCCGTACTCGGTCCCGTACTCGGT-3' and the 10 bp satellite oligo probe 5'-AATAACATAGAATAACATAGAATAACATAGAATAACATAGAATAACATAG-3' were chemically synthesized and labeled at the 5' end with Cy3 or at the 3' end with fluorescein (New England Biolabs). DNA probes derived from clones or PCR products were labeled by nick translation with digoxygenin-11-dUTP (Roche) using the DIG-Nick Translation Mix (Roche). Digoxigenin labeled probes were detected with Anti-Digoxigenin-Rhodamine, Fab fragments (Roche) in a 1:200 dilution, following supplier recommendations. Digital images were obtained using a Zeiss Axiover 200 microscope equipped with a cooled Charge-Coupled Device camera. The fluorescent signals were recorded separately as grey-scale digital images and then pseudo-colored and merged using Adobe Photoshop software.

Immunofluorescence-FISH on extended chromatin fibers

Extended chromatin fibers were prepared from S2 cells by centrifuging 5×10^4 cells onto slides at 800 rpm for 4 min in a Cytospin 4 (Thermo Shandon, Pittsburgh, PA), and then slides were dipped into salt detergent lysis buffer (25 mM Tris, pH 7.5, 500 mM NaCl, and 1% Triton X-100) for 25 minutes, slowly and steadily removed using an in-house made device consisting of a modified EasyDip™ Slide Staining System connected to a peristaltic pump, and subsequently fixed in 4% paraformaldehyde (PFA) for 5 minutes. Slides were incubated in 1× PBST (1× PBS ± 0.05% Tween-20) for 15 minutes. Slides were dipped again in the former lysis dilution for 15 minutes, after which they were slowly and steadily removed. Slides were blocked in 1× PBS, 0.1% Triton X-100, 4%

formaldehyde for 10 minutes at room temperature and washed for 5 minutes in 1x PBS before proceeding to immunostaining. Slides were blocked in goat serum (Zymed Laboratories) for 30 minutes and incubated overnight at 4ºC with a chicken anti-CID antibody (Blower and Karpen 2001), diluted to 1:100 in blocking buffer. Slides were washed 3 times for 5 minutes in 1x PBST and incubated for 1 hour at 37ºC in Alexa 488 anti-chicken secondary antibody (Molecular probes). Slides were then washed 3 times in 1x PBST and 3 times in 1x PBS. After immunofluorescence with CID antibodies, slides were re-fixed in 4% formaldehyde for 15 minutes and then hybridized to the dodeca probe. For each slide, around 250ng of dodeca probe were precipitated with 3M Sodium Acetate and absolute ethanol, re-suspended in hybridization solution (50% formamide, 10% dextran sulfate, 2x SSC) and denatured for 10 min at 80ºC. Slides were incubated at 37ºC for 24 hr.

DNA sample preparation for NMR and MS experiments

Oligonucleotides were purchased from Integrated DNA Technologies, IDT, Coralville, IA, USA. Samples for NMR experiments were dissolved in 9:1 H_2O/D_2O. Buffer conditions: 25 mM sodium phosphate, 100 mM NaCl pH 4.0 for C rich sequences and 25 mM potassium phosphate, 100 mM KCl pH 7.0 for G rich sequences. The latter were previously annealed by heating at 90 ºC for 5 minutes and cooling down to room temperature overnight.

Samples for MS experiments were dissolved at 100 µM in 100 mM NH_4OAc buffer at pH 7 and pH 4. pH was adjusted by adding acetic acid and NH_3 aliquots.

NMR experiments

All NMR spectra were acquired in Bruker spectrometers operating at 600 and 800 MHz, equipped with cryoprobes and processed with the TOPSPIN software. A jump-and-return pulse sequence [45] was employed to observe the rapidly exchanging protons in 1D H_2O experiments. In most of the experiments in H_2O, water suppression was achieved by including a WATERGATE module in the pulse sequence prior to acquisition.

Circular Dichroism spectroscopy

Circular dichroism spectra at different temperatures were recorded on a Jasco J-810 spectropolarimeter fitted with a thermostated cell holder. CD spectra were recorded in 25 mM sodium phosphate buffer, pH 4, with 100 mM NaCl (100 µM oligo concentration). CD melting curves were recorded at the wavelength of the larger positive band, 285 nm with a heating rate of 0.5 $ºC.min^{-1}$.

Mass spectrometry

All ESI-MS experiments were carried out in the negative ion mode on an Exactive ESI-Orbitrap mass spectrometer (Thermo Scientific, Bremen, Germany). The ESI spray voltage and capillary voltage used were -2.75 kV and -20 V, respectively. The capillary temperature was set to 150 ºC.

Tube lens and skimmer voltage were fixed to 180 V and -10 V, respectively. Samples were injected at a flow rate of 4 µL min^{-1}.

References

1 Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098-1102 (2001).

2 Abad, J. P. & Villasante, A. Searching for a common centromeric structural motif: *Drosophila* centromeric satellite DNAs show propensity to form telomeric-like unusual DNA structures. *Genetica* **109**, 71-75 (2000).

3 Malik, H. S. & Henikoff, S. Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. *Genetics* **157**, 1293-1298 (2001).

4 Schueler, M. G., Swanson, W., Thomas, P. J. & Green, E. D. Adaptive evolution of foundation kinetochore proteins in primates. *Mol. Biol. Evol.* **27**, 1585-1597 (2010).

5 Müller, S. *et al.* Phosphorylation and DNA Binding of HJURP Determine its centromeric recruitment and function in $CenH3^{CENP-A}$ loading. *Cell reports* **8**, 190-203 (2014).

6 Kato, T. *et al.* Activation of Holliday Junction–Recognizing Protein involved in the chromosomal stability and immortality of cancer cells. *Cancer Res.* **67**, 8544-8553 (2007).

7 du Sart, D. *et al.* A functional neo-centromere formed through activation of a latent human centromere and consisting of non-alpha-satellite DNA. *Nat. Genet.* **16**, 144-153 (1997).

8 Williams, B. C., Murphy, T. D., Goldberg, M. L. & Karpen, G. H. Neocentromere activity of structurally acentric mini-chromosomes in *Drosophila*. *Nat. Genet.* **18**, 30-38 (1998).

9 Rocchi, M., Archidiacono, N., Schempp, W., Capozzi, O. & Stanyon, R. Centromere repositioning in mammals. *Heredity* **108**, 59-67 (2011).

10 Han, F., Gao, Z. & Birchler, J. A. Reactivation of an inactive centromere reveals epigenetic and structural components for centromere specification in maize. *The Plant Cell Online* **21**, 1929-1939 (2009).

11 Agudo, M. *et al.* A dicentric chromosome of *Drosophila melanogaster* showing alternate centromere inactivation. *Chromosoma* **109**, 190-196 (2000).

12 Sullivan, B. A. & Willard, H. F. Stable dicentric X chromosomes with two functional centromeres. *Nat. Genet.* **20**, 227-228 (1998).

13 Fisher, A. M. *et al.* Centromeric inactivation in a dicentric human Y; 21 translocation chromosome. *Chromosoma* **106**, 199-206 (1997).

14 Steiner, N. C. & Clarke, L. A novel epigenetic effect can alter centromere function in fission yeast. *Cell* **79**, 865-874 (1994).

15 Hayden, K. E. *et al.* Sequences associated with centromere competency in the human genome. *Mol. Cell. Biol.* **33**, 763-772 (2013).

16 Allshire, R. C. & Karpen, G. H. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nat. Rev. Genet.* **9**, 923-937 (2008).

17 Fachinetti, D. *et al.* A two-step mechanism for epigenetic specification of centromere identity and function. *Nat. Cell Biol.* **15**, 1056-1066 (2013).

18 Black, B. E. & Cleveland, D. W. Epigenetic centromere propagation and the nature of CENP-a nucleosomes. *Cell* **144**, 471-479 (2011).

19 Blower, M. D., Sullivan, B. A. & Karpen, G. H. Conserved organization of centromeric chromatin in flies and humans. *Dev. Cell* **2**, 319-330 (2002).

20 Ribeiro, S. A. *et al.* A super-resolution map of the vertebrate kinetochore. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10484-10489 (2010).

21 Sullivan, B. A. & Karpen, G. H. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nat. Struct. Mol. Biol.* **11**, 1076-1083 (2004).

22 Abad, J. *et al.* Pericentromeric regions containing 1.688 satellite DNA sequences show anti-kinetochore antibody staining in prometaphase chromosomes of *Drosophila melanogaster*. *Molecular & general genetics: MGG* **264**, 371-377 (2000).

23 Abad, J. P. *et al.* Dodeca satellite: a conserved G+ C-rich satellite from the centromeric heterochromatin of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 4663-4667 (1992).

24 Agudo, M. *et al.* Centromeres from telomeres? The centromeric region of the Y chromosome of *Drosophila melanogaster* contains a tandem array of telomeric HeT-A-and TART-related sequences. *Nucleic Acids Res.* **27**, 3318-3324 (1999).

25 Lohe, A. R., Hilliker, A. & Roberts, P. Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. *Genetics* **134**, 1149 (1993).

26 Méndez-Lago, M. *et al.* Novel sequencing strategy for repetitive DNA in a *Drosophila* BAC clone reveals that the centromeric region of the Y chromosome evolved from a telomere. *Nucleic Acids Res.*, **37**, 2264-2273 (2009).

27 Sun, D. *et al.* Inhibition of human telomerase by a G-quadruplex-interactive compound. *J. Med. Chem.* **40**, 2113-2116 (1997).

28 Losada, A., Abad, J., Agudo, M. & Villasante, A. Long-range analysis of the centromeric region of *Drosophila melanogaster* chromosome 3. *Chromosome Res.* **8**, 651-653 (2000).

29 Carmena, M., Abad, J. P., Villasante, A. & Gonzalez, C. The *Drosophila melanogaster* dodecasatellite sequence is closely linked to the centromere and can form connections between sister chromatids during mitosis. *J. Cell Sci.* **105**, 41-50 (1993).

30 Ferrer, N., Azorin, F., Villasante, A., Gutierrez, C. & Abad, J. Centromeric dodeca-satellite DNA sequences form fold-back structures. *J. Mol. Biol.* **245**, 8-21 (1995).

31 Gallego, J., Chou, S.-H. & Reid, B. R. Centromeric pyrimidine strands fold into an intercalated motif by forming a double hairpin with a novel T:G:G:T tetrad: solution structure of the d(TCCCGTTTCCA) dimer. *J. Mol. Biol.* **273**, 840-856 (1997).

32 Gehring, K., Leroy, J.-L. & Guéron, M. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature* **363**, 561-565 (1993).

33 Leroy, J. L., Gehring, K., Kettani, A. & Gueron, M. Acid multimers of oligodeoxycytidine strands: stoichiometry, base-pair characterization, and proton exchange properties. *Biochemistry* **32**, 6019-6031 (1993).

34 Lieblein, A. L., Kramer, M., Dreuw, A., Furtig, B. & Schwalbe, H. The nature of hydrogen bonds in cytidine...H+...cytidine DNA base pairs. *Angew. Chem. Int. Ed. Engl.* **51**, 4067-4070 (2012).

35 Guo, K. *et al.* Formation of pseudosymmetrical G-quadruplex and i-motif structures in the proximal promoter region of the RET oncogene. *J. Am. Chem. Soc.* **129**, 10220-10228 (2007).

36 Cui, J., Waltman, P., Le, V. H. & Lewis, E. A. The effect of molecular crowding on the stability of human c-MYC promoter sequence i-motif at neutral pH. *Molecules* **18**, 12751-12767 (2013).

37 Sun, D. & Hurley, L. H. The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression. *J. Med. Chem.* **52**, 2863-2874 (2009).

38 Gatti, M. & Pimpinelli, S. Functional elements in *Drosophila melanogaster* heterochromatin. *Annu. Rev. Genet.* **26**, 239-276 (1992).

39 Southern, E., Anand, R., Brown, W. & Fletcher, D. A model for the separation of large DNA molecules by crossed field gel electrophoresis. *Nucleic Acids Res.* **15**, 5925-5943 (1987).

40 Losada, A., Abad, J. P., Agudo, M. & Villasante, A. The analysis of Circe, an LTR retrotransposon of *Drosophila melanogaster*, suggests that an insertion of non-LTR retrotransposons into LTR elements can create chimeric retroelements. *Mol. Biol. Evol.* **16**, 1341-1346 (1999).

41 Török, T., Gorjánácz, M., Bryant, P. J. & Kiss, I. Prod is a novel DNA-binding protein that binds to the 1.686 g/cm3 10 bp satellite repeat of *Drosophila melanogaster*. *Nucleic Acids Res.* **28**, 3551-3557 (2000).

42 Blower, M. D. & Karpen, G. H. The role of *Drosophila* CID in kinetochore formation, cell-cycle progression and heterochromatin interactions. *Nat. Cell Biol.* **3**, 730-739 (2001).

43 Andreyeva, E. N. *et al.* High-resolution analysis of *Drosophila* heterochromatin organization using SuUR Su (var) 3-9 double mutants. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 12819-12824 (2007).

44 Käs, E. & Laemmli, U. K. *In vivo* topoisomerase II cleavage of the *Drosophila* histone and satellite III repeats: DNA sequence and structural characteristics. *The EMBO journal* **11**, 705 (1992).

45 Plateau, P. & Gueron, M. Exchangeable proton NMR without base-line distorsion, using new strong-pulse sequences. *J. Am. Chem. Soc.* **104**, 7310-7311 (1982).

Acknowledgements

We gratefully acknowledge D. V. Laurents for revision of the manuscript. We thank G. H. Karpen for instructing M. Mendez-Lago in the preparation of extended chromatin fibers and for critical comments on the manuscript. We thank the Structural Biophysical Chemistry platform of the IECB (CNRS UMS3033 / Inserm US001) for the access to the mass spectrometry facility and Dr. F. Rosu for his kind assistance. We also thank A. Losada for the initial long-range restriction analysis. We acknowledge financial support from MICINN (CTQ2010-21567-C02-02 to C.G.; BFU2011-30295-C02-01 to A.V.), the Inserm (ATIP-Avenir Grant no. R12086GS to V.G.), the Conseil Régional Aquitaine (Grant no. 20121304005 to V.G.), the EU (FP7-PEOPLE-2012-CIG-333611 to V.G.), the Wellcome Trus, and the institutional grant from the Fundación Ramón Areces to the Centro de Biología Molecular “Severo Ochoa”. M.G. was supported by the FPI-fellowship BES-2009-027909.

SUPPLEMENTARY DATA FOR

The structure of an endogenous *Drosophila* centromere reveals the prevalence of tandemly repeated sequences able to form i-motifs.

Miguel Garavís[1,2], María Méndez-Lago[1,6], Valérie Gabelica[3,4], Siobhan L. Whitehead[5], Carlos González[2] * and Alfredo Villasante[1] *

[1]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Nicolás Cabrera 1, 28049 Madrid, Spain. [2]Instituto de Química Física Rocasolano, CSIC, Serrano 119, 28006 Madrid, Spain. [3]Univ. Bordeaux, ARNA Laboratory, IECB, 2 rue Robert Escarpit F-33600 Pessac, France. [4] Inserm ARNA Laboratory, 146 rue Leo Saignat, F-33000 Bordeaux, France. [5]The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, United Kingdom.

[6]Present address: Centro Nacional de Análisis Genómico, Baldiri Reixac 4, 08028 Barcelona, Spain.

Supplementary Figures

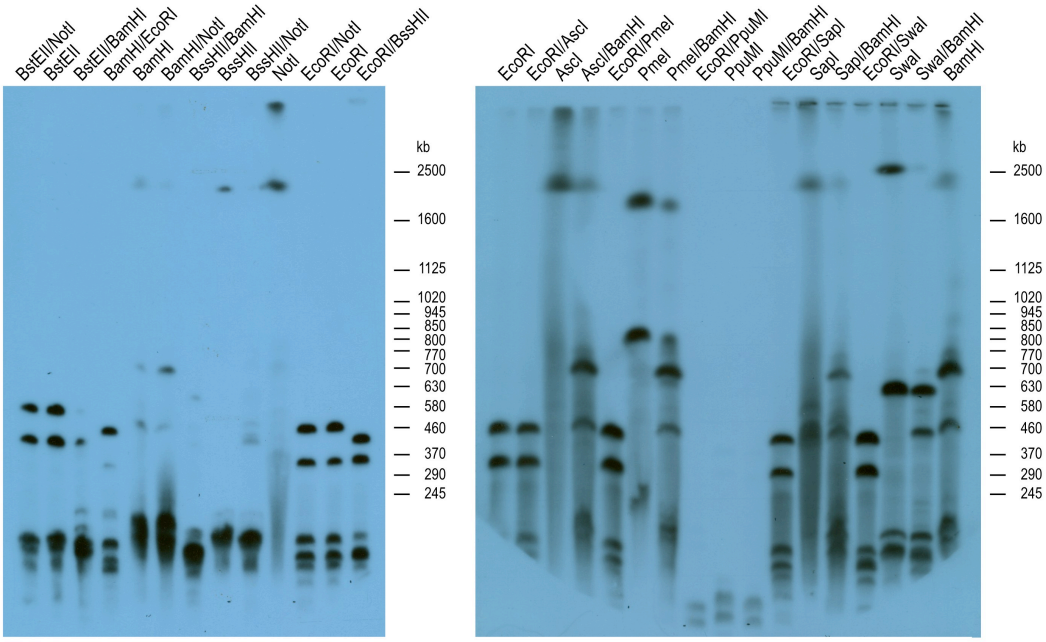


Figure S1. *Pulsed-field gel electrophoresis analysis of the dodeca satellite DNA. High molecular embryos was digested with one or two restriction enzymes, fractionated (using a "Waltzer" apparatus) on a 1.2% (w/v) agarose gel run at 150 V for 28 h with a pulse time of 80 s, blotted and hybridized with the dodeca satellite probe pBK6E218. Two representative gels are shown.*

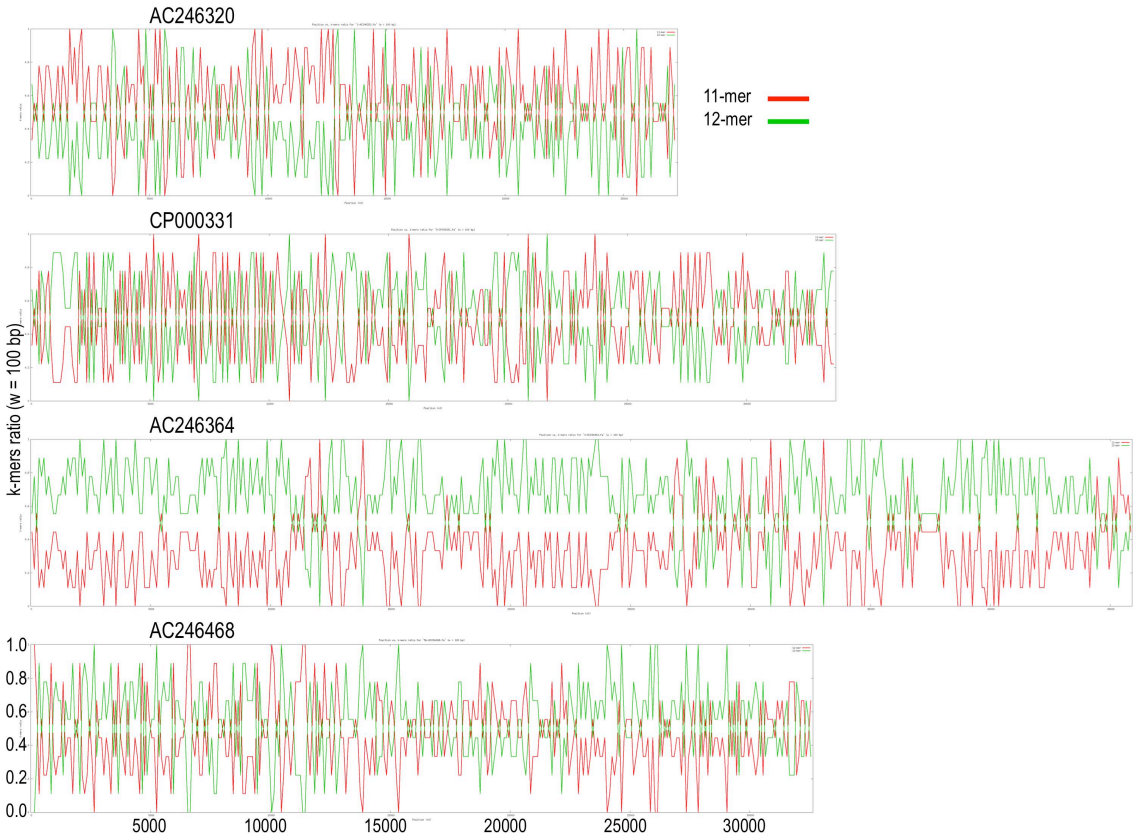


Figure S2. *Linear representation of undeca (in red) and dodeca (in green) repeats observed in two scaffolds from the dodeca satellite block I (AC246320 and CP00031), and two scaffolds from the dodeca satellite block II (AC246364 and AC246468).*

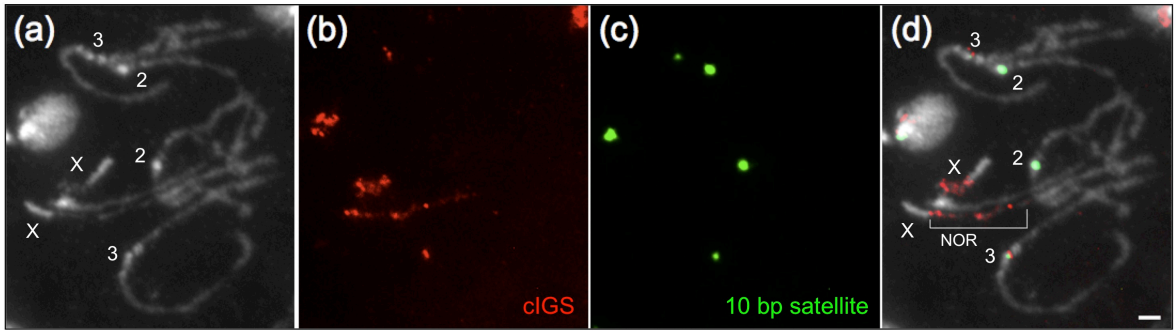


Figure S3. *Fluorescence in situ hybridization of cIGS and 10 bp satellite to prometaphase chromosomes from Oregon R females. The hybridization was performed under low-stringency conditions (25°C). (a) Prometaphase chromosomes counterstained with DAPI. (b) Hybridization signals from a cIGS probe (in red). (c) Hybridization signals from a 10 bp satellite probe (in green). (d) DAPI stained superimposed with hybridization signals. The nucleolus organizer region (NOR) is indicated. The Scale bar is 2 µm.*

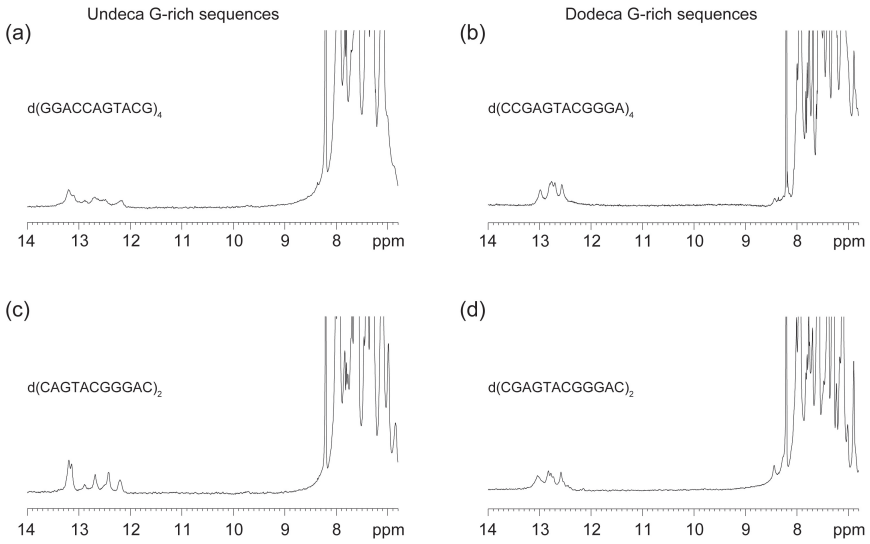


Figure S4. *Imino region of the NMR spectra of the G-rich stands. Top: oligos containing four repeats of the undeca sequences (a) and dodeca sequences (b). Bottom: Oligos containing two repeats of undeca (c) and dodeca (d). Experimental conditions:* [46] *= 0.8 mM, 25 mM potassium phosphate, T= 5ºC, pH 7.*

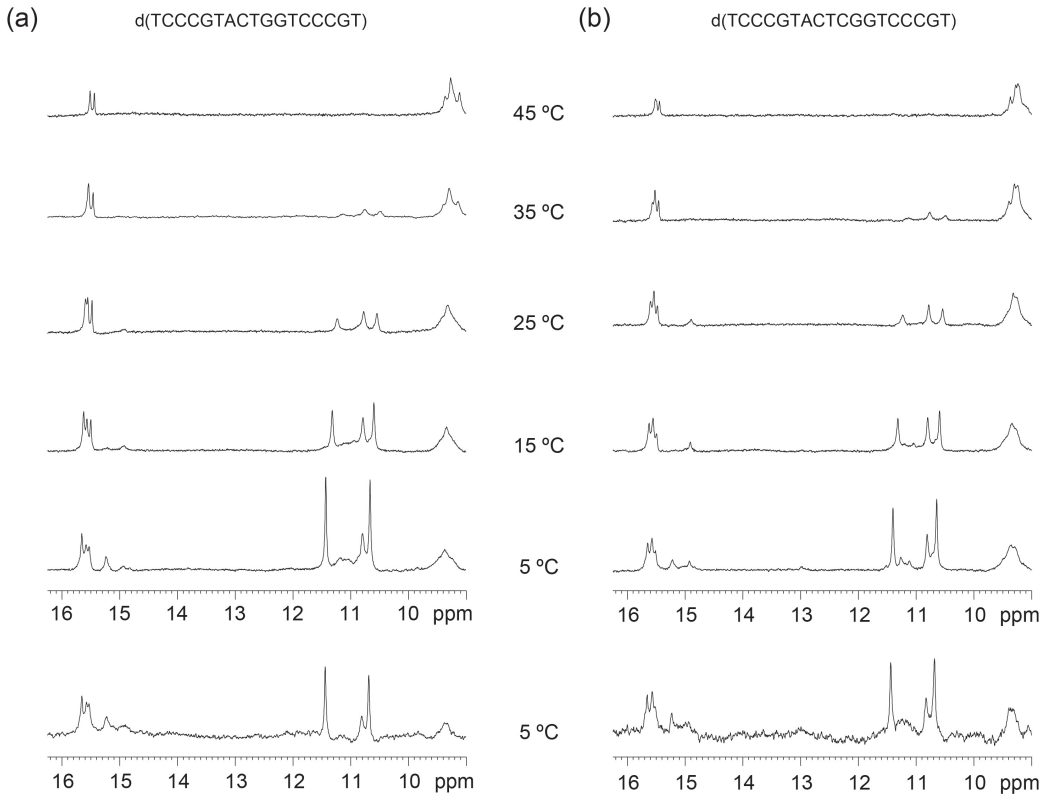


Figure S5. *Imino region of the NMR spectra of the C-rich strands of the undeca (a) and dodeca (b) at different temperatures. Experimental conditions: [oligo]=0.8 mM, 25 mM sodium phosphate, pH 4. Bottom spectra: 100 mM NH_4OAc [oligo]=100 µM, T= 5ºC, pH 4.*

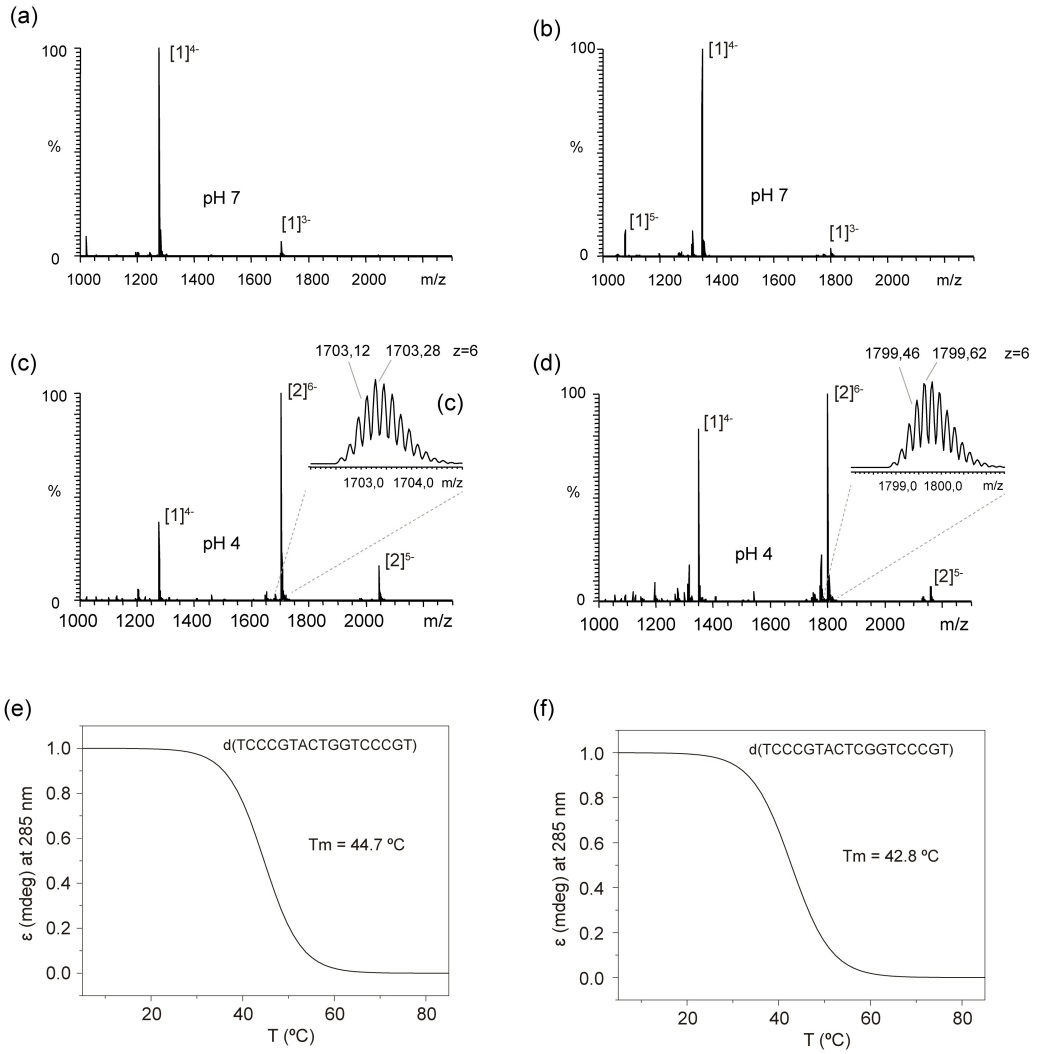


Figure S6. *Mass spectrometry data of undeca (left panels) and dodeca (right panels). Buffer conditions: 100 mM NH_4OAc, pH 7 (a, b); and 100 mM NH_4AcO, pH 4 (c, d). Zoom views of the dimer peaks showing the isotopic distribution are displayed in the insets of the panels e and f (separation between two consecutive ^{13}C isotopes in the isotopic distribution corresponds to m/z equals to 1/z, consequently in main isotopic distribution the z value is 6, and the mass is that of a dimer). Normalized CD melting curves of the undeca (e) and dodeca (f). Same conditions as Figure 3.*

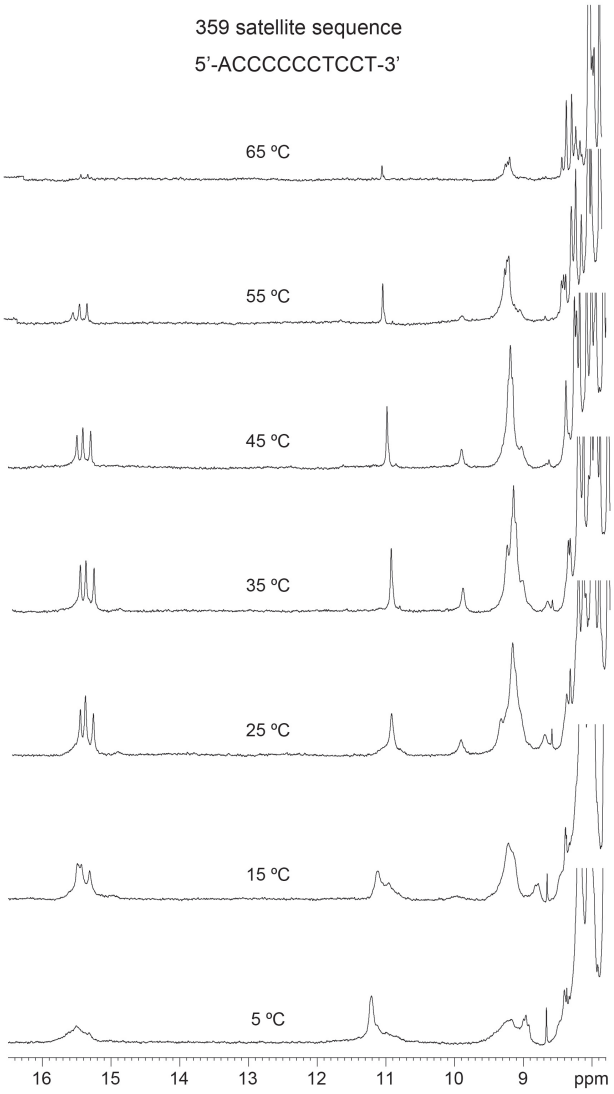


Figure S7. *Imino region of the NMR spectra (at different temperatures) of the C-rich region found at the 359 satellite DNA. The spectra show sharp imino signals around 15-16 ppm, which are characteristic features of i-motif formation. Buffer conditions: 25 mM NaPi, 100 mM NaCl pH=4.0. Oligo concentration: [359] = 0.6 mM*

Discusión

DISCUSIÓN

Los telómeros y los centrómeros son regiones heterocromáticas de los cromosomas eucariotas cuyas funciones respectivas son la protección del final del cromosoma y la segregación del material genético durante la división celular. Se ha propuesto que el origen evolutivo de estas regiones del cromosoma tuvo como punto de partida la invasión masiva del genoma de una arqueobacteria hospedadora con elementos móviles procedentes de una α-proteobacteria (simbionte). En esta tesis se plantea que estos elementos móviles fueron empleados en mecanismos de reparación de rotura de DNA, consistentes en la inserción de estas secuencias transponibles en los sitios de ruptura. Nuestra hipótesis sugiere que, en algún momento de la evolución, la incorporación en tándem de elementos transponibles, con una distribución sesgada en nucleótidos de guanina y citosina (G/C) en los sitios de ruptura, no condujo a la reparación del DNA, sino a la formación de las primeras estructuras con función telomérica. Además, proponemos que el sesgo en contenido en G/C pudo ser seleccionado por la capacidad que tienen estos nucleótidos de formar algún tipo de estructura no canónica. Postulamos también que, los primeros telómeros, o prototelómeros, debían tener una función doble que asegurase tanto la protección, como la correcta segregación cromosómica. Más adelante se separarían ambas funciones, con la aparición del centrómero en las regiones subteloméricas.

El posible papel de las estructuras no canónicas de DNA en el origen y evolución de los telómeros se discute en base a los datos estructurales referentes a los distintos tipos de telómeros que se conocen. Por un lado, la composición de las secuencias teloméricas sintetizadas por la enzima telomerasa, muestran un claro desequilibrio en residuos G/C entre las hebras que las componen. El caso más significativo es el de la secuencia $d(TTAGGG)_n$, que se considera el motivo de repetición ancestral de los organismos eucariotas y que ha mostrado ser capaz de formar cuádruplex de guanina *in vitro* y de hacerlo con mayor facilidad que ninguna otra secuencia telomérica estudiada. El análisis de las secuencias teloméricas de organismos con una menor actividad telomerasa y con un menor contenido en guaninas, revela que éstas están compuestas por secuencias cortas repetidas en tándem y por retroelementos que conservan un sesgo de nucleótidos G/C entre sus hebras. En estos organismos, la inserción de retrotransposones non-LTR complementa la baja actividad telomerasa para lograr el mantenimiento de la longitud de los telómeros. Es importante resaltar que la inserción orientada de estos elementos mantiene la abundancia de guaninas en una de las hebras frente a la complementaria y esto ha llevado a sugerir que los cuádruplex de guanina podrían tener un papel en la función telomérica. Por último, los telómeros de los organismos que carecen de telomerasa, como la mosca *Drosophila melanogaster,* se mantienen mediante la incorporación en tándem de elementos transponibles y mediante fenómenos de recombinación. De nuevo, en este tipo de telómeros el contenido en nucleótidos guanina y citosina es desigual entre ambas hebras pues la adición en tándem de los retroelementos mantiene la asimetría de éstos al final del cromosoma. Este caso nos recuerda al mecanismo ancestral propuesto para el mantenimiento de los primeros telómeros. El mecanismo de formación de los telómeros de *D. melanogaster*, hace que las secuencias terminales sufran constantes cambios y por consiguiente que no haya una secuencia telomérica definida. Sin embargo, las proteínas teloméricas de estos organismos son capaces de reconocer dichas secuencias, lo que sugiere que las proteínas teloméricas

reconocen motivos de estructura secundaria y no de estructura primaria. En otros organismos carentes de telomerasa, como es el caso de algunas levaduras, el mantenimiento de los telómeros se consigue a través de la amplificación y el reordenamiento de secuencias subteloméricas o de rDNA. Se ha comprobado que las secuencias subteloméricas de estas levaduras mantienen un desequilibrio de G/C entre hebras que, por consiguiente, puede ocurrir también en los telómeros de estos organismos. Por otro lado, se ha visto que la proteína de unión a DNA de cadena sencilla, POT1 (siglas del inglés *Protection Of Telomeres* 1), reconoce los extremos 3' monohebra de los rDNA de forma independiente de secuencia, lo que sugiere, de nuevo, la formación de determinadas estructuras secundarias en los telómeros.

La obtención de información sobre la estructura de secuencias teloméricas y centroméricas es esencial para la verificación de esta hipótesis y por ello ha sido uno de los objetivos principales de esta tesis.

El estudio estructural de moléculas de RNA telomérico de un número elevado de repeticiones de la secuencia telomérica humana, se llevó a cabo utilizando técnicas biofísicas y de molécula individual. Este estudio fue posible gracias al diseño de un sistema que permitiera producir tanto moléculas de TERRA a gran escala - para ser analizadas mediante RMN y CD- como ensamblajes con las características necesarias para ser estudiadas mediante pinzas ópticas. La combinación de ambos tipos de técnicas permitió la obtención de información complementaria sobre la estructura de TERRA.

A partir de datos de RMN y CD se ha podido determinar que una molécula de RNA telomérico formada por 16 repeticiones del motivo r(GGGUUA) ($TERRA_{16}$) se pliega formando cuatro cuádruplex paralelos, esto es, la estructura presenta el mayor número de cuádruplex que se pueden formar con la cantidad de residuos de guanina que contiene la secuencia. Estos resultados concuerdan con la inmensa mayoría de los datos publicados sobre la estructura de cuádruplex de RNA telomérico, que indican que la conformación paralela es la predilecta en estas moléculas. Se han reportado, sin embargo, algunos casos excepcionales, como el de las secuencias de RNA telomérico de 22 y 45 nucleótidos, que en presencia de cationes amonio, presentan una banda en torno a 300 nm en su espectro de dicroísmo circular. Esta banda es característica de la conformación antiparalela y se atribuye, en estos casos, a la posible existencia de una fracción de moléculas que se pliegan formando dicha conformación o a la formación de ordenamientos particulares de los *loops* UUA. Otra evidencia similar de la posible existencia de conformaciones antiparalelas en moléculas de RNA telomérico se observó en el espectro de RMN de una secuencia de 23 nucleótidos, cuya región imino presenta señales que apuntan a un equilibrio entre conformaciones. Sin embargo, en el caso de $TERRA_{16}$, los datos de dicroísmo circular indican claramente que los cuádruplex que se forman adoptan una conformación paralela.

Las medidas de temperatura de desnaturalización de $TERRA_{16}$ confirman que, tal y como indican trabajos previos, la estabilidad térmica de estas moléculas no se correlaciona con la longitud de la secuencia. Por otro lado, el aumento de temperatura de desnaturalización en presencia de mayores concentraciones de iones K^+ corrobora el efecto estabilizador que tienen estos cationes sobre los cuádruplex de DNA y RNA.

Los espectros de RMN de las moléculas de TERRA analizadas, muestran las señales características de los protones imino que forman parte de las tétradas de guanina. Además, en algunos espectros se observan

señales a desplazamientos químicos más altos, típicos de formación de pares A:U, y que podrían ser consecuencia de la existencia de interacciones entre nucleótidos de los *loops*. Este tipo de interacciones han sido observadas previamente en la estructura de un cuádruplex de guanina de RNA telomérico obtenida por difracción de rayos X. En ese caso, la secuencia r(^{Br}UAGGGUUAGGGU) forma un cuádruplex de guanina dimérico y paralelo. En el cristal se observan interacciones entre los nucleótidos uracilo y adenina del extremo 5' de un cuádruplex, con los mismos nucleótidos del cuádruplex contiguo en la red cristalina. Por otro lado, los nucleótidos UUA de uno de los *loops* de la estructura, se sitúan apilados entre sí. En el caso de las moléculas de TERRA estudiadas en esta tesis, las interacciones uracilo-adenina (U:A) podrían darse también entre nucleótidos de un mismo *loop* o entre nucleótidos de diferentes *loops*. Los contactos entre *loops* diferentes, implicarían la interacción entre dos unidades de cuádruplex de guanina, que podría darse de forma lateral (a través de las caras de los *loops propeller*) o por apilamiento de una unidad sobre la siguiente. En este último caso, el *loop* que une un cuádruplex con el siguiente, es el que establecería interacciones con los *loops* laterales de los cuádruplex de guanina apilados. Para determinar el origen concreto de estas interacciones sería preciso contar con información estructural de mayor resolución.

Para llevar a cabo los experimentos de fuerza-extensión con pinzas ópticas fue necesario desarrollar un sistema de estudio adecuado para su uso en esta técnica. El protocolo desarrollado para la preparación de los ensamblajes moleculares empleados en los estudios de TERRA en pinzas ópticas, ha demostrado ser válido, y puede ser aplicado para la realización de estudios posteriores de las propiedades mecánicas de moléculas de TERRA o de otras moléculas de RNA.

Las curvas de fuerza-extensión obtenidas para una secuencia de TERRA formada por 5 repeticiones r(GGGUUA), indicaron que la fuerza necesaria para desplegar un cuádruplex de guanina de RNA es significativamente menor que la fuerza que se precisa para desplegar un cuádruplex de guanina de DNA. Esta diferencia puede atribuirse a una menor intensidad de las interacciones de apilamiento en el cuádruplex de RNA. Esta situación sería análoga a la que se da en los dúplex tipo A de RNA y tipo B de DNA, donde se pudo determinar que la doble cadena de RNA muestra un menor valor del módulo de estiramiento que la doble cadena equivalente de DNA. Por otro lado, la menor fuerza de ruptura del cuádruplex de guanina de RNA es consistente con las medidas realizadas sobre cuádruplex de guanina formados por la secuencia ILPR (del inglés *Insuline-Linked Polymorphic Region*) del gen de la insulina, que indican que los cuádruplex paralelos presentan fuerzas de apertura menores que los antiparalelos. Estos resultados, sin embargo, contrastan con los obtenidos en el grupo del Prof. Hanbin Mao, donde se llevó a cabo el estudio de una molécula de TERRA formada por 4 repeticiones de la secuencia telomérica humana. En dicho trabajo, el análisis de las curvas de fuerza-extensión revela la presencia de dos poblaciones con longitudes de contorno diferentes. Una de las poblaciones se atribuye a una especie parcialmente plegada, que se asigna a un posible intermedio en el equilibrio entre las estructuras plegada y desplegada. La otra población se asigna a la especie completamente plegada. Los valores de fuerza de ruptura para ambas poblaciones son mayores que los obtenidos en nuestro trabajo e implican una mayor estabilidad mecánica del cuádruplex de RNA frente al de DNA. Las diferencias en los valores obtenidos en ambos trabajos pueden ser atribuidas a diversas causas. Por un lado, la molécula

objeto de estudio es diferente en cada caso, teniendo una de ellas una repetición más de la secuencia telomérica, lo cual podría influir en la estabilidad mecánica de la molécula. Por otro lado, el aparato de pinzas ópticas utilizado para realizar los experimentos no es el mismo en los dos casos. En las curvas de fuerza-extensión que se muestran en esta tesis, se han considerado como control de la correcta calibración del aparato, los valores de fuerza de sobreestiramiento de los híbridos de DNA:RNA que sirven como asas para generar la fuerza de estiramiento. Los valores de sobreestiramiento medidos son del orden de los característicos para doble cadena de DNA y de RNA. En el caso del trabajo realizado en el grupo del Prof. Mao, no se detalla cómo se verifica la correcta calibración del aparato. En cuanto a la diferencia en el número de poblaciones observadas; esta divergencia podría ser atribuida a que el número de medidas realizadas en nuestro trabajo no es lo suficientemente alto como para resolver la existencia de dos poblaciones diferentes.

Los experimentos de estiramiento de secuencias de TERRA con 16 y 25 repeticiones del motivo r(GGGUUA) reflejaron situaciones variadas. En numerosos casos, la apertura de los sucesivos cuádruplex se produce a diferentes valores de fuerza que, según el caso, son más o menos distantes entre sí. Además, las fuerzas de apertura correspondientes a dos ciclos de estiramiento-relajación realizados sobre la misma molécula no son iguales. Esto indica que los cuádruplex de guanina son capaces de replegarse durante la etapa de relajación y que las fuerzas de ruptura de la molécula replegada y la original no tienen porqué ser las mismas. En algunos casos, las curvas de fuerza-extensión mostraron la apertura de dos cuádruplex de guanina al mismo valor de fuerza. Estos fenómenos de apertura múltiple serían altamente improbables de no existir interacción entre las dos unidades que se despliegan de forma simultánea, lo que sugiere que en algunas de las moléculas, las unidades de cuádruplex consecutivas interaccionan entre sí. Estos resultados abren la puerta a futuros estudios acerca del comportamiento mecánico del RNA telomérico. La optimización de este sistema de análisis puede ser de gran utilidad en el estudio de interacciones de ligandos que estabilizan cuádruplex de guanina de TERRA.

La utilización de ligandos que unen y estabilizan cuádruplex de guanina teloméricos, se ha convertido en una atractiva estrategia terapéutica contra el cáncer. La estabilización de estas estructuras conduce a la disrupción de la heterocromatina telomérica y en consecuencia a la senescencia de la célula maligna. En principio, esta estrategia es válida para cualquier célula cancerosa, afectando tanto a las células que mantienen la longitud de sus telómeros mediante la acción de la enzima telomerasa como a aquellas que usan el mecanismo alternativo de elongación de telómeros denominado ALT.

La gran mayoría de los compuestos que unen y estabilizan cuádruplex de guanina, se basan en plataformas aromáticas, tales como acridinas, porfirinas o antraquinonas, que reconocen y estabilizan el cuádruplex mediante apilamiento sobre las tétradas de guanina. Estos compuestos no cumplen las características recomendables para ser utilizados como fármacos debido, en general, a su excesiva masa molecular y a que la abundancia de grupos aromáticos en su estructura dificulta su entrada en las células. Por otro lado, los métodos que permiten encontrar nuevos compuestos que interaccionen con cuádruplex son escasos, de manera que la mayoría de ligandos de cuádruplex que se han descubierto hasta ahora son el

resultado de aplicar un diseño racional sobre estructuras que mostraron afinidad por otras conformaciones de DNA. En esta tesis se ha puesto a punto un nuevo método para la identificación de compuestos capaces de unir cuádruplex de RNA. Para ello se han combinado el uso de RMN basada en detección de flúor con las estrategias de cribado, denominadas genéricamente como FBDD (siglas del inglés *Fragment-Based Drug Discovery*), basadas en la búsqueda de compuestos pequeños (o fragmentos) que interaccionen con la diana de interés.

La ^{19}F-RMN es una herramienta con numerosas ventajas para su empleo en FBDD. Por un lado, permite detectar interacciones débiles entre dos moléculas. Además, el efecto de cambio en anchura de línea, que supone la base para la detección de interacciones, es más acentuado sobre señales de flúor que sobre señales de protón. Por último, las señales de flúor son finas y el intervalo de desplazamientos químicos en el que pueden aparecer es muy amplio, de manera que es posible adquirir espectros de muestras que contienen varios compuestos fluorados sin que en ellos haya solapamiento de señales. Esta ventaja es fundamental a la hora de realizar cribados de compuestos, ya que permite agrupar los compuestos de la librería y registrar la respuesta de todos ellos en el mismo espectro, lo cual reduce, tanto el tiempo de operación, como la cantidad de molécula diana que se emplea en el proceso.

La metodología que se presenta en esta tesis para la búsqueda de pequeños ligandos fluorados de TERRA, se fundamenta en el hecho de que la anchura de línea de las señales de RMN es inversamente proporcional a la constante de relajación transversal R_2, y ésta, a su vez, varía en función de la masa molecular, siendo mayor en compuestos grandes que en compuestos pequeños. La constante de relajación R_2 del ligando, se ve alterada de manera transitoria debido al tiempo que permanece unido a la molécula más grande con la que interacciona. Por lo tanto, aquellos compuestos que interaccionan con la molécula diana experimentan una modificación parcial de sus constantes de relajación, comportándose transitoriamente como moléculas de mucho mayor tamaño y en consecuencia produciendo un ensanchamiento en sus señales de RMN. Este efecto es perceptible a concentraciones muy bajas de ligando unido y en equilibrios de asociación con constantes de afinidad débiles, lo cual permite realizar los experimentos usando cantidades pequeñas de molécula diana y detectar compuestos que interaccionen con baja afinidad. Por otro lado, la detección del ensanchamiento de las señales de flúor del ligando, permite visualizar el efecto de manera muy clara, sin la interferencia de señales correspondientes a la molécula diana, a los disolventes o a los búferes de la muestra.

La molécula utilizada para la implementación de este método fue el RNA telomérico $TERRA_{16}$ que, tal y como se ha descrito anteriormente, forma cuatro unidades consecutivas de cuádruplex de guanina paralelos. La elección de esta molécula como diana se basa, por un lado, en que su longitud se aproxima a la de la región repetida de los TERRA presentes en la célula y, por otro lado, en que la existencia de varios cuádruplex en la misma estructura, puede proporcionar sitios de unión específicos que no son posibles en sistemas que formen un solo cuádruplex.

La aplicación del método de búsqueda de ligandos resultó satisfactoria, obteniéndose una tasa de positivos del 5.6%, lo que confirma que la metodología utilizada es aplicable a moléculas de ácidos nucleícos.

Los 20 compuestos identificados como ligandos de $TERRA_{16}$ muestran una importante heterogeneidad estructural, presentando motivos estructurales de los cuales se desconocía su afinidad por estructuras tipo cuádruplex. Estos resultados ponen de relieve el potencial de esta metodología para aumentar la variabilidad química a considerar en el diseño y síntesis de nuevos ligandos de cuádruplex de guanina.

Varios de los ligandos obtenidos en el proceso de búsqueda se seleccionaron para ser validados y para evaluar su afinidad y selectividad frente a otras conformaciones de RNA y DNA. La elección de los compuestos se hizo atendiendo a criterios de disponibilidad y de solubilidad, eligiéndose aquellos cuya solubilidad estimada en agua fuera mayor.

La validación de varios de los ligandos encontrados se llevo a cabo desde varias perspectivas. Por un lado, se utilizó de nuevo ^{19}F-RMN, aplicando el mismo principio utilizado en el proceso de cribado pero en muestras que contenían tan sólo el compuesto que se pretendía validar y en una concentración mayor que la usada en las muestras mezcla. Además, utilizando las mismas muestras empleadas en la validación por ^{19}F-RMN, se realizaron experimentos basados en ^{1}H-RMN como el STD (del inglés *Saturation Transfer Difference*). La aplicación del experimento STD a este sistema de estudio resulta muy apropiada, por cumplir éste los requisitos necesarios de cantidades bajas de molécula diana, gran exceso de ligando en la muestra y diferencia de tamaño notable entre molécula diana y ligando. Los resultados obtenidos en los experimentos de ^{19}F-RMN y STD confirmaron la validez de los compuestos seleccionados como ligandos de $TERRA_{16}$.

Con el fin de obtener más información acerca del modo de interacción de los ligandos, se examinó el efecto que producían los compuestos sobre un cuádruplex de guanina dimérico y paralelo de TERRA, cuya estructura tridimensional ha sido determinada, tanto por RMN, como por cristalografía de Rayos X. El menor tamaño de esta estructura permite obtener espectros de ^{1}H-RMN con señales finas y con buena resolución, lo que permite analizar los cambios que se producen en las señales de TERRA y de los ligandos, a diferentes proporciones de ambos componentes. De este modo fue posible calcular las constantes de disociación (K_D) de varios de los ligandos. Los valores de K_D obtenidos están dentro de un rango de afinidad moderada, lo cual es esperado para ligandos resultantes de cribados basados en pequeños compuestos. Un nivel de afinidad débil es, además, conveniente para la realización de los experimentos de cribado, ya que favorece un régimen de intercambio rápido, lo cual es requisito indispensable para detectar el efecto de ensanchamiento sobre las señales de flúor de los ligandos.

Los cambios producidos en la región imino de la molécula de TERRA en presencia de los distintos ligandos indican que cada uno de ellos afecta de forma diferente a la estructura del cuádruplex. Estas diferencias podrían ser consecuencia de modos de interacción que implican a distintas regiones del cuádruplex, aunque sería necesario un análisis más profundo de la estructura de los complejos TERRA:ligando para confirmarlo.

La selectividad por la estructura cuádruplex de los ligandos encontrados es una variable a tener en cuenta a la hora de desarrollar moléculas con un objetivo terapéutico. En este trabajo se estudió la capacidad de algunos de los ligandos encontrados, de interaccionar con estructuras de DNA y RNA frecuentes en la célula. Recurriendo de nuevo a los experimentos de ^{19}F-RMN descritos anteriormente, se evaluó si los

ligandos seleccionados interaccionaban con una molécula de tRNA. Los resultados mostraron que tan solo uno de los ligandos sufría el efecto de ensanchamiento de su señal de flúor en presencia de tRNA. Para evaluar la selectividad frente a B-DNA, se utilizaron experimentos monodimensionales de ^{1}H-RMN, registrándose los cambios observados en la región imino del espectro de un dúplex de ocho pares de bases, en presencia de ligando. Solo 2 de los 7 ligandos que se testaron, produjeron cambios destacables en el espectro de RMN. El conjunto de los resultados indica que algunos de los compuestos solo interaccionan con estructuras tipo cuádruplex de guanina, lo que unido a los distintos modos de unión que parecen presentar, hace que estos compuestos reúnan características muy apropiadas para ser usados como compuestos de partida en el desarrollo de ligandos con mayor afinidad y selectividad.

Los cuádruplex de guanina formados por secuencias de DNA telomérico presentan una importante variabilidad conformacional. Un ejemplo de ello es el dodecámero d(TAGGGTTAGGGT) que en disolución da lugar a un equilibrio entre la conformación paralela y la antiparalela, lo que origina un espectro de RMN con un gran número de señales en la región imino, correspondientes a ambas conformaciones. En presencia de los compuestos resultantes del cribado, la secuencia de DNA telomérico presenta un espectro de RMN cuya región imino es mucho más sencilla y que se corresponde con la conformación paralela. Esta observación indica que los compuestos seleccionados desplazan el equilibrio hacia la conformación paralela lo cual es razonable si se tiene en cuenta que los ligandos se han encontrado utilizando como molécula diana un RNA telomérico, cuya conformación predilecta es la paralela. La obtención de estructuras tridimensionales de los complejos formados por esta secuencia de DNA y los ligandos, podría revelar las razones por las que estos compuestos favorecen la conformación paralela. Dada la conocida heterogeneidad conformacional que presentan las estructuras cuádruplex en general y el DNA telomérico en particular, estos datos estructurales serían de utilidad para el diseño de ligandos dirigidos a estabilizar una conformación determinada.

El DNA telomérico presenta una región monohebra rica en guaninas y una región de doble cadena en la que la hebra rica en guaninas se encuentra unida a su hebra complementaria rica en citosinas (ver sección 3.1 de la Introducción). Así como la hebra rica en guaninas es capaz de plegarse adoptando estructuras tipo cuádruplex de guanina, la hebra rica en citosinas, cuando no está hibridada con su complementaria, también puede adoptar estructuras no canónicas basadas en el apilamiento de pares citosina:citosina (C:C^{+}), denominadas *i-motif* (del inglés: *intercalated motif*). Al contrario de lo que ocurre con los cuádruplex de guanina, que precisan secuencias con un número considerable de guaninas para formarse, los *i-motifs* se pueden formar a partir de secuencias con mucha menor presencia de citosinas. Tal afirmación queda patente en la publicación de Escaja et al. 2012 (ver Anexo) en la que se muestra la formación de un *i-motif* dimérico a partir de una secuencia de DNA que presenta tan solo dos citosinas en su estructura primaria. Este hecho nos planteó la posibilidad de explorar la posible formación de estructuras tipo *i-motif* en secuencias centroméricas, en las cuales, la formación de cuádruplex de guanina está mucho más restringida debido a la ausencia de tramos de guanina tan largos y numerosos como los que aparecen en las secuencias teloméricas. Por otro lado, los estudios de Gallego et al. sobre la secuencia CENP-B box del DNA alfoide

humano, que habían revelado que la hebra rica en citosinas de dicha secuencia forma un *i-motif* dimérico, nos invitaron a investigar la capacidad que tienen otras secuencias de DNA centromérico, de plegarse formando estructuras tipo *i-motif*.

El DNA centromérico está compuesto por secuencias altamente repetidas denominadas DNA satélite. Es conveniente señalar que las secuencias que forman parte de los centrómeros de cualquier organismo no se conocen en su totalidad, debido a las dificultades que entraña tanto la obtención de mapas físicos, como el ensamblaje de los largos tramos de secuencias repetidas presentes en la región centromérica. En esta tesis se ha llevado a cabo el estudio de secuencias derivadas del centrómero humano, uno de los centrómeros de los que más información se conoce. Concretamente, presentamos el estudio estructural de la secuencia A-box que, al igual que la secuencia CENP-B box, está formada por 17 pb y forma parte del satélite alfoide, el satélite centromérico presente en los cromosomas humanos. La secuencia A-box aparece en algunas de las unidades de repetición del satélite alfoide en las que no está presente la secuencia CENP-B box y es la única secuencia rica en guaninas y citosinas que aparece en el cromosoma Y, cuyo centrómero carece de la secuencia CENP-B box.

Existen dos variantes de la secuencia A-box, las cuales se diferencian entre sí en un solo nucleótido (timina o citosina en la posición 10). Los espectro de RMN de las secuencias de 17 nucleótidos correspondientes a las hebras ricas en guanina de ambas variantes (no mostrados en esta tesis) muestran que dichas secuencias no forman estructuras tipo cuádruplex. Por el contrario, el análisis de RMN de las secuencias ricas en citosina de las dos variantes, revela que ambas secuencias forman estructuras tipo *i-motif* a pH ácido. Para facilitar la determinación de la estructura tridimensional de la variante con citosina en la posición 10 (HS), se estudiaron secuencias truncadas de la misma, que en un primer análisis por RMN, mostraron las mismas señales en la región imino que la secuencia completa, lo que implica que los apareamientos que estabilizan la estructura son los mismos en las secuencias truncadas y en la completa.

La determinación de la molecularidad de estas estructuras es un paso clave para la resolución de la estructura tridimensional de las mismas. La espectrometría de masas con ionización por electrospray (ESI-MS) es una técnica de gran utilidad para este fin, ya que las condiciones de ionización son tales que no destruyen las interacciones no covalentes que mantienen unidas unas moléculas con otras. Los experimentos de ESI-MS muestran que las estructuras formadas por estas secuencias a pH ácido son dímeros. Las condiciones de trabajo necesarias en espectrometría de masas hacen imposible obtener espectros en presencia de iones Na^+. En su defecto, lo más habitual es utilizar sales de amonio, que se asemeja en tamaño y tiene la misma carga que el ión Na^+. En este trabajo se ha comprobado que los *i-motifs* formados en presencia de iones amonio conservan la misma estructura que en presencia de iones Na^+. Posteriormente, se pudo comprobar que los oligonucleótidos disueltos en H_2O en ausencia de sales de amonio y sodio también eran capaces de adoptar las mismas estructuras. En este aspecto, cabe destacar que la influencia de los iones presentes en la muestra -ya sea procedentes del búffer o de sales- en la estabilidad de la estructura, no es comparable al efecto estabilizador que producen, tanto concentraciones altas de oligo, como determinados valores de pH.

Los espectros de RMN bidimensionales permitieron determinar la estructura tridimensional de una secuencia truncada de 10 nucleótidos derivada de la secuencia de HS. La estructura es un *i-motif* dimérico en el que dos moléculas de DNA se asocian con orientación cabeza-cola a través de la formación de cuatro pares $C:C^+$ flanqueados por dos pares T:T y otros dos pares más externos A:T Hoogsteen. La orientación de las hebras que se observa en esta estructura contrasta con la orientación cabeza-cabeza del *i-motif* que forma la secuencia CENP-B box. La composición de nucleótidos de los *loops* tiene una gran influencia en la orientación de las hebras en los *i-motifs* diméricos, ya que éstos pueden dar lugar a interacciones que estabilicen una u otra conformación. Así, por ejemplo, en la secuencia de CENP-B box, la orientación cabeza-cabeza permite que los nucleótidos de los *loops* formen una tétrada G:T:G:T que se dispone al final de la serie de pares $C:C^+$ intercalados. El mismo tipo de tétrada se observa flanqueando ambos extremos del *i-motif* dimérico formado por la secuencia d(TCGTTTCGTT) (ver Anexo). Otro motivo de interacción entre los nucleótidos de los *loops* es la formación de pares T:T que se ha observado en otros *i-motifs* diméricos de secuencias no centroméricas. Los pares T:T son bastante similares a los pares $C:C^+$ lo que les permite formar parte de las estructuras *i-motif* mediante su intercalación entre pares $C:C^+$ o como pares terminales de la estructura.

En la estructura de la secuencia A-box se observan dos pares A:T tipo Hoogsteen que se disponen en los extremos de la estructura. La formación del par tiene lugar entre la adenina terminal de la secuencia y una de las cuatro timinas que componen el *loop*. Este tipo de apareamiento se ha observado con anterioridad en el *i-motif* tetramérico formado por la secuencia d(TAACCC) del telómero humano que fue determinada mediante difracción de Rayos X. En el caso de la estructura de la A-box, el empleo de RMN ha permitido estudiar la influencia del pH en la formación del par A:T Hoogsteen. La obtención de espectros a distintos valores de pH reveló que el par A:T se forma a valores de pH inferiores a 4, lo cual indica que la protonación de la adenina puede ser el factor determinante para que se produzca el apareamiento. Esto implica que el pK_a de la adenina que intervienen en el par (pK_a 4.1), ha aumentado en relación al valor que toma como nucleótido aislado (pK_a 3.5). El aumento del valor de pK_a de adeninas ha sido observado previamente en otras estructuras, particularmente en adeninas que intervienen en apareamientos no Watson-Crick. La modificación del pK_a de la adenina es consecuencia de que la protonación de la adenina puede suponer un aumento de la estabilidad de la estructura debido a varias posibles causas. Por un lado, la protonación del nitrógeno N1, aumenta la acidez de los protones amino de la adenina, lo que a su vez aumenta la fortaleza del par de bases. Por otro lado, el enlace carbono-nitrógeno amino adquiere un carácter parcial de doble enlace si el N1 está protonado, lo que impide parcialmente la rotación del enlace y en consecuencia puede favorecer la formación del par de bases. Además, la protonación de la adenina puede favorecer interacciones con las cargas negativas del esqueleto azúcar-fosfato.

Una característica común en las estructuras *i-motif* es la formación de cuatro surcos en la estructura, dos de ellos mucho más anchos que los otros dos. Los surcos estrechos o surcos menores de los *i-motif* diméricos pueden separar hebras pertenecientes a la misma cadena o a cadenas diferentes. En el caso de la estructura de la A-box, el surco menor se forma entre hebras de la misma cadena, es decir, es

intramolecular. Lo mismo ocurre con la mayoría de los *i-motif* diméricos que se han resuelto hasta el momento. En el caso del *i-motif* formado por la secuencia d(mCCTTTACC) el surco menor separa hebras de cadenas diferentes. Sería necesario profundizar en el estudio de este tipo de estructuras para establecer cuáles son los aspectos estructurales que determinan la naturaleza intramolecular o intermolecular de los surcos.

Un análisis más exhaustivo de la estructura de la variante de A-box con timina en lugar de citosina en la posición 10 (HST) reveló que esta secuencia forma un *i-motif* estabilizado igualmente por cuatro pares $C:C^+$ que, en este caso, se forman entre de residuos de citosina más alejados en la secuencia. Esta reorganización de apareamientos de la secuencia indica que las dos variantes de A-box dan lugar a estructuras similares pese a la diferencia en sus estructuras primarias. En el caso de la estructura de HST, los espectros de RMN bidimensionales muestran que los pares $C:C^+$ están flanqueados por cuatro pares de bases A:T de los cuales los dos más internos son Watson-Crick reverso y los dos más externos son tipo Hoogsteen. Por lo tanto, ambas estructuras están estabilizadas por ocho pares de bases no canónicos.

Al igual que ocurre en humanos, el centrómero de los autosomas del ratón común ó *Mus musculus* está formado por el DNA satélite "minor" y éste presenta la secuencia CENP-B box mientras que el DNA satélite del centrómero del Y (que muestra homología con el satélite "minor") no la tiene y en su lugar aparece una secuencia rica en guaninas y citosinas. Con el fin de continuar la exploración de la capacidad de ciertas secuencias centroméricas de formar *i-motif*, se adquirieron espectros de RMN de esta secuencia a pH ácido e igualmente se detectó la formación de estructuras basadas en la formación de pares $C:C^+$.

La posibilidad de formar estructuras tipo *i-motif* por parte de secuencias centroméricas también se exploró en secuencias derivadas de DNA satélites simples. Como representante de este tipo de DNA satélite, se eligió el DNA satélite dodeca, presente en el centrómero del cromosoma 3 de la mosca *D. melanogaster*. El DNA satélite dodeca esta compuesto por la repetición en tándem de una secuencia de 11 o 12 pares de bases que muestra un predominio de nucleótidos de guanina en una de las hebras y de citosinas en su complementaria. La estructura del centrómero del cromosoma 3 muestra que el satélite dodeca se distribuye en dos grandes bloques que tienen un contenido diferente de unidades de 11 y 12 pares de bases.

El estudio estructural de las secuencias derivadas de dodeca se llevo a cabo estudiando oligonucleótidos correspondientes a ambas hebras del satélite dodeca. Tanto las secuencias ricas en guanina como las ricas en citosina fueron elegidas de manera que permitieran la formación de cuádruplex de guanina e *i-motif*, respectivamente. Los espectros de RMN de las secuencias ricas en guanina indicaban que éstas no forman cuádruplex de guanina y sí estructuras basadas en apareamientos tipo Watson-Crick. Este resultado confirma lo observado por Ferrer, et al. quienes determinaron que secuencias derivadas de la misma región, forman horquillas muy estables de DNA. Por otro lado, el estudio por RMN de las secuencias ricas en citosina revela que dichas secuencias se pliegan formando estructuras tipo *i-motif* a pH ácido. Además, se pudo determinar mediante espectrometría de masas que los *i-motif* formados son diméricos.

Los datos recabados durante el estudio estructural de secuencias centroméricas llevado a cabo en esta tesis sugieren que la estructura *i-motif* puede tener un papel en la formación del centrómero. Así,

planteamos que el *i-motif* puede actuar como motivo de unión entre hebras de DNA distantes en la secuencia primaria. En la heterocromatina centromérica se distinguen nucleosomas que contienen la proteína CENP-A, que aporta la marca epigenética a los centrómeros, y nucleosomas canónicos que contienen la histona H3. Ambos tipos de nucleosomas aparecen intercalados a lo largo del centrómero. Además, se sabe que los nucleosomas con CENP-A interaccionan lateralmente entre sí y originan una lámina en la superficie de la constricción primaria del cromosoma. El DNA centromérico se enrolla en torno a los nucleosomas quedando las secuencias CENP-B box y A-box del satélite alfoide, situadas a la salida y a la entrada de los nucleosomas. En el caso de la secuencia CENP-B box, se sabe que la unión de la proteína CENP-B en nucleosomas con CENP-A, no impide la digestión del DNA unido, de manera que la secuencia CENP-B box permanece accesible. El ordenamiento resultante de la interacción lateral de los nucleosomas centroméricos junto con la accesibilidad de las secuencias CENP-B box y A-box hacen posible que se establezcan interacciones DNA-DNA entre secuencias distantes en secuencia primaria. La formación de estructuras tipo *i-motif* serviría como motivo de autorreconocimiento entre las secuencias de DNA. En este aspecto es importante destacar que la naturaleza bimolecular de las estructuras estudiadas ejemplifica el tipo de interacción que puede tener lugar de acuerdo al modelo propuesto en este trabajo. Las observaciones hechas hasta el momento indican que este modelo es compatible con las características de otros centrómeros bastante bien caracterizados como es el caso de los de organismos *Mus musculus* y *D. melanogaster*.

Para la formación de las estructuras *i-motif* es necesario producir la protonación de las citosinas que forman parte de los pares $C:C^+$ lo cual tiene lugar en condiciones de pH ácido. Por este motivo, el papel del *i-motif* como estructura relevante en el contexto celular siempre se ha puesto en entredicho. Sin embargo, estudios recientes han demostrado que determinadas secuencias son capaces de formar *i-motif* en condiciones de pH neutro. Además, se ha observado que en presencia de moléculas que imitan la densidad molecular propia del ambiente celular, se favorece la formación de *i-motifs* a pH neutro. Así mismo, la superhelicidad negativa que se produce en el DNA de doble cadena durante la transcripción, también posibilita la formación de estas estructuras a pH fisiológico. El descubrimiento de la estructura *i-motif in vivo* se presenta como un paso fundamental para garantizar la viabilidad de esta estructura en el contexto celular y para entender su posible relevancia biológica.

La existencia de un motivo estructural del DNA característico del centrómero, explicaría lo que se conoce como "la paradoja del centrómero" que plantea el hecho aparentemente contradictorio de que las proteínas centroméricas presenten un elevado grado de conservación mientras que las secuencias de DNA centromérico sufren una rápida evolución como consecuencia de los fenómenos de recombinación que acontecen con gran frecuencia entre secuencias repetidas. El reconocimiento por parte de las proteínas centroméricas, de una estructura de DNA y no de unas secuencias determinadas, daría explicación a este hecho.

El advenimiento de nuevas técnicas de secuenciación, que permitan conocer en su totalidad la secuencia de DNA de los centrómeros, junto con el desarrollo de nuevas investigaciones destinadas a

revelar la estructura de la heterocromatina centromérica resultarán claves para la verificación de las hipótesis expuestas en esta tesis.

Conclusiones

CONCLUSIONES

1. La molécula de RNA telomérico formada por 16 repeticiones de la secuencia telomérica humana r(GGGUUA) adopta una estructura formada por cuatro cuádruplex de guaninas paralelos que interaccionan entre sí.
2. Las moléculas de RNA telomérico formadas por uno o varios cuádruplex de guaninas, se despliegan de manera no reversible al someterse a una fuerza de estiramiento. Esta fuerza, del orden de pN, es significativamente menor que la necesaria para desplegar moléculas de DNA telomérico.
3. Se ha implementado con éxito una metodología basada en la utilización de ^{19}F-RMN para la búsqueda de nuevos compuestos de bajo peso molecular con afinidad por moléculas de RNA. Mediante el cribado de una pequeña librería de 355 compuestos fluorados, se han identificado 20 de ellos como ligandos de RNA telomérico (tasa de éxito del 5.6%).
4. El estudio de las interacciones de algunos de los nuevos ligandos descubiertos, con moléculas de RNA telomérico, reveló que todos ellos provocan una estabilización del cuádruplex de guaninas paralelo y que el grado de afinidad de los compuestos se encuentra en el rango 100-1000 μM.
5. El análisis de la interacción de estos compuestos con otras moléculas de RNA y DNA ha permitido identificar varios compuestos que interaccionan selectivamente con cuádruplex de guaninas paralelos.
6. Se ha determinado la estructura tridimensional a pH ácido, de la A-box del DNA satélite alfoide humano. La estructura es un dímero simétrico en el que las dos hebras se asocian con orientación cabeza-cola. Los 10 primeros nucleótidos están involucrados en el motivo de auto-reconocimiento, mientras que el segmento 3'-terminal se encuentra desordenado. La estructura es un *i-motif* cuyo núcleo está constituido por cuatro pares de bases C:C$^+$ y dos pares T:T, todos ellos formados por nucleótidos de hebras paralelas, flanqueado por dos pares de bases A:T tipo Hoogsteen.
7. El estudio de la variante principal de la A-box humana (sustitución C9→T9) indica que el motivo estructural se mantiene, aunque los nucleótidos involucrados en el mismo son distintos. La estructura de esta variante contiene lazos más largos e interacciones tipo reverse Watson-Crick en lugar de los pares T:T, lo que probablemente resulta menos eficiente para estabilizar el dímero.
8. El estudio estructural de secuencias derivadas de la hebra rica en citosinas del DNA satélite dodeca, presente en el centrómero del cromosoma 3 de *Drosophila melanogaster,* muestra que dichas secuencias son capaces de formar estructuras diméricas tipo *i-motif*.
9. La observación recurrente de estructuras no-canónicas (*i-motif* y cuádruplex de guaninas) en estudios *in vitro* de ácidos nucleicos centroméricos y teloméricos, sugiere que su posible papel en la estructura y la evolución de centrómeros y telómeros no es una hipótesis descabellada y merece ser considerada en futuros estudios.

Bibliografía

BIBLIOGRAFÍA

Abrescia, N.G., González, C., Gouyette, C., and Subirana, J.A. (2004). X-ray and NMR studies of the DNA oligomer d(ATATAT): Hoogsteen base pairing in duplex DNA. Biochemistry *43*, 4092-4100.

Altona, C., and Sundaralingam, M. (1973). Conformational analysis of the sugar ring in nucleosides and nucleotides. Improved method for the interpretation of proton magnetic resonance coupling constants. J. Am. Chem. Soc. *95*, 2333-2344.

Altona, C.t., and Sundaralingam, M. (1972). Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. J. Am. Chem. Soc. *94*, 8205-8212.

Amad, M.a.H., Cech, N.B., Jackson, G.S., and Enke, C.G. (2000). Importance of gas-phase proton affinities in determining the electrospray ionization response for analytes and solvents. J. Mass Spectrom. *35*, 784-789.

Arias-Gonzalez, J.R. (2013). Optical Tweezers to Study Viruses. In Structure and Physics of Viruses (Springer), pp. 273-304.

Arias-Gonzalez, J.R. (2014). Single-molecule portrait of DNA and RNA double helices. Integr. Biol. *6*, 904-925.

Arora, R., Brun, C.M., and Azzalin, C.M. (2012). Transcription regulates telomere dynamics in human cancer cells. RNA *18*, 684-693.

Azzalin, C.M., Reichenbach, P., Khoriauli, L., Giulotto, E., and Lingner, J. (2007). Telomeric repeat–containing RNA and RNA surveillance factors at mammalian chromosome ends. Science *318*, 798-801.

Bacolla, A., and Wells, R.D. (2009). Non-B DNA conformations as determinants of mutagenesis and human disease. Mol. Carcinog. *48*, 273-285.

Bah, A., Wischnewski, H., Shchepachev, V., and Azzalin, C.M. (2012). The telomeric transcriptome of Schizosaccharomyces pombe. Nucleic Acids Res. *40*, 2995-3005.

Beck, J.L., Colgrave, M.L., Ralph, S.F., and Sheil, M.M. (2001). Electrospray ionization mass spectrometry of oligonucleotide complexes with drugs, metals, and proteins. Mass Spectrom. Rev. *20*, 61-87.

Biffi, G., Tannahill, D., McCafferty, J., and Balasubramanian, S. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. Nat. Chem. *5*, 182-186.

Blackburn, G.M. (2006). Nucleic acids in chemistry and biology (Royal Society of Chemistry).

Borgias, B.A., and James, T.L. (1990). MARDIGRAS-A procedure for matrix analysis of relaxation for discerning geometry of an aqueous structure. J. Magn. Reson. (1969) *87*, 475-487.

Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res. *34*, 5402-5415.

Burger, A.M., Dai, F., Schultes, C.M., Reszka, A.P., Moore, M.J., Double, J.A., and Neidle, S. (2005). The G-quadruplex-interactive molecule BRACO-19 inhibits tumor growth, consistent with telomere targeting and interference with telomerase function. Cancer Res. *65*, 1489-1496.

Canalia, M., and Leroy, J.-L. (2009). [5mCCTCTCTCC]4: An i-motif tetramer with intercalated T*T pairs. J. Am. Chem. Soc. *131*, 12870-12871.

Canalia, M., and Leroy, J.L. (2005). Structure, internal motions and association–dissociation kinetics of the i-motif dimer of d(5mCCTCACTCC). Nucleic Acids Res. *33*, 5471-5481.

Case, D.A., Pearlman, D.A., Caldwell, J.W., Cheatham III, T.E., Wang, J., Ross, W.S., Simmerling, C., Darden, T., Merz, K.M., and Stanton, R.V. (2002). AMBER 7. University of California, San Francisco.

Chang, C.-C., Chien, C.-W., Lin, Y.-H., Kang, C.-C., and Chang, T.-C. (2007). Investigation of spectral conversion of d(TTAGGG)4 and d(TTAGGG)13 upon potassium titration by a G-quadruplex recognizer BMVC molecule. Nucleic Acids Res. *35*, 2846-2860.

Choo, K.H. (2001). Domain organization at the centromere and neocentromere. Dev. Cell *1*, 165-177.

Cole, R.B. (2000). Some tenets pertaining to electrospray ionization mass spectrometry. J. Mass Spectrom. *35*, 763-772.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. *117*, 5179-5197.

Cusanelli, E., Romero, C.A.P., and Chartrand, P. (2013). Telomeric noncoding RNA TERRA is induced by telomere shortening to nucleate telomerase molecules at short telomeres. Mol. Cell *51*, 780-791.

Davis, J.T. (2004). G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry. Angew. Chem. Int. Ed. *43*, 668-698.

Day, H.A., Huguin, C., and Waller, Z.A. (2013). Silver cations fold i-motif at neutral pH. Chem. Commun. *49*, 7696-7698.

Donohue, J., and Trueblood, K.N. (1960). Base pairing in DNA. J. Mol. Biol. *2*, 363-371.

Drygin, D., Siddiqui-Jain, A., O'Brien, S., Schwaebe, M., Lin, A., Bliesath, J., Ho, C.B., Proffitt, C., Trent, K., and Whitten, J.P. (2009). Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis. Cancer Res. *69*, 7653-7661.

Episkopou, H., Draskovic, I., Van Beneden, A., Tilman, G., Mattiussi, M., Gobin, M., Arnoult, N., Londoño-Vallejo, A., and Decottignies, A. (2014). Alternative Lengthening of Telomeres is characterized by reduced compaction of telomeric chromatin. Nucleic Acids Res. *42*, 4391-4405.

Escaja, N., Viladoms, J., Garavís, M., Villasante, A., Pedroso, E., and González, C. (2012). A minimal i-motif stabilized by minor groove G:T:G:T tetrads. Nucleic Acids Res. *40*, 11737-11747.

Flynn, R.L., Centore, R.C., O'Sullivan, R.J., Rai, R., Tse, A., Songyang, Z., Chang, S., Karlseder, J., and Zou, L. (2011). TERRA and hnRNPA1 orchestrate an RPA-to-POT1 switch on telomeric single-stranded DNA. Nature *471*, 532-536.

Franks, F. (1975). The hydrophobic interaction. In Water A comprehensive treatise (Springer), pp. 1-94.

Gallego, J., Chou, S.-H., and Reid, B.R. (1997). Centromeric pyrimidine strands fold into an intercalated motif by forming a double hairpin with a novel T:G:G:T tetrad: solution structure of the d(TCCCGTTTCCA) dimer. J. Mol. Biol. *273*, 840-856.

Gallego, J., Golden, E.B., Stanley, D.E., and Reid, B.R. (1999). The folding of centromeric DNA strands into intercalated structures: a physicochemical and computational study. J. Mol. Biol. *285*, 1039-1052.

Grand, C.L., Han, H., Munoz, R.M., Weitman, S., Von Hoff, D.D., Hurley, L.H., and Bearss, D.J. (2002). The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth *in vivo*. Mol Cancer Ther *1*, 565-573.

Greenwood, J., and Cooper, J.P. (2012). Non-coding telomeric and subtelomeric transcripts are differentially regulated by telomeric and heterochromatin assembly factors in fission yeast. Nucleic Acids Res. *40*, 2956-2963.

Harrington, J.J., Van Bokkelen, G., Mays, R.W., Gustashaw, K., and Willard, H.F. (1997). Formation of *de novo* centromeres and construction of first-generation human artificial microchromosomes. Nat. Genet. *15*, 345-355.

Haschemeyer, A., and Rich, A. (1967). Nucleoside conformations: an analysis of steric barriers to rotation about the glycosidic bond. J. Mol. Biol. *27*, 369-384.

Hofstadler, S.A., and Griffey, R.H. (2001). Analysis of noncovalent complexes of DNA and RNA by mass spectrometry. Chem. Rev. *101*, 377-390.

Horvath, M.P., and Schultz, S.C. (2001). DNA G-quartets in a 1.86 Å resolution structure of an *Oxytricha nova* telomeric protein-DNA complex. J. Mol. Biol. *310*, 367-377.

Ikeno, M., Grimes, B., Okazaki, T., Nakano, M., Saitoh, K., Hoshino, H., McGill, N.I., Cooke, H., and Masumoto, H. (1998). Construction of YAC–based mammalian artificial chromosomes. Nat. Biotechnol. *16*, 431-439.

Incles, C.M., Schultes, C.M., Kempski, H., Koehler, H., Kelland, L.R., and Neidle, S. (2004). A G-quadruplex telomere targeting agent produces p16-associated senescence and chromosomal fusions in human prostate cancer cells. Mol. Cancer Ther. *3*, 1201-1206.

Kebarle, P. (2000). A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. J. Mass Spectrom. *35*, 804-817.

Leonetti, C., Scarsella, M., Riggio, G., Rizzo, A., Salvati, E., D'Incalci, M., Staszewsky, L., Frapolli, R., Stevens, M.F., and Stoppacciaro, A. (2008). G-quadruplex ligand RHPS4 potentiates the antitumor activity of camptothecins in preclinical models of solid tumors. Clin. Cancer Res. *14*, 7284-7291.

Lin, C.-T., Tseng, T.-Y., Wang, Z.-F., and Chang, T.-C. (2011). Structural conversion of intramolecular and intermolecular G-quadruplexes of bcl2mid: the effect of potassium concentration and ion exchange. J. Phys. Chem. B. *115*, 2360-2370.

Loo, J.A. (2000). Electrospray ionization mass spectrometry: a technology for studying noncovalent macromolecular complexes. Int. J. Mass Spectrom. *200*, 175-186.

Luke, B., Panza, A., Redon, S., Iglesias, N., Li, Z., and Lingner, J. (2008). The Rat1p 5' to 3' exonuclease degrades telomeric repeat-containing RNA and promotes telomere elongation in *Saccharomyces cerevisiae*. Mol. Cell *32*, 465-477.

Mergny, J.-L., Lacroix, L., Han, X., Leroy, J.-L., and Helene, C. (1995). Intramolecular folding of pyrimidine oligodeoxynucleotides into an i-DNA motif. J. Am. Chem. Soc. *117*, 8887-8898.

Nergadze, S.G., Farnung, B.O., Wischnewski, H., Khoriauli, L., Vitelli, V., Chawla, R., Giulotto, E., and Azzalin, C.M. (2009). CpG-island promoters drive transcription of human telomeres. RNA *15*, 2186-2194.

Ng, L.J., Cropley, J.E., Pickett, H.A., Reddel, R.R., and Suter, C.M. (2009). Telomerase activity is associated with an increase in DNA methylation at the proximal subtelomere and a reduction in telomeric transcription. Nucleic Acids Res. *37*, 1152-1159.

Nikolova, E.N., Kim, E., Wise, A.A., O'Brien, P.J., Andricioaei, I., and Al-Hashimi, H.M. (2011). Transient Hoogsteen base pairs in canonical duplex DNA. Nature *470*, 498-502.

Olaussen, K.A., Dubrana, K., Domont, J., Spano, J.-P., Sabatier, L., and Soria, J.-C. (2006). Telomeres and telomerase as targets for anticancer drug development. Crit. Rev. Oncol. Hematol. *57*, 191-214.

Pennarun, G., Granotier, C., Gauthier, L.R., Gomez, D., Hoffschir, F., Mandine, E., Riou, J.-F., Mergny, J.-L., Mailliet, P., and Boussin, F.D. (2005). Apoptosis related to telomere instability and cell cycle alterations in human glioma cells treated by new highly selective G-quadruplex ligands. Oncogene *24*, 2917-2928.

Pfeiffer, V., and Lingner, J. (2012). TERRA promotes telomere shortening through exonuclease 1–mediated resection of chromosome ends. PLoS Genet. *8*, e1002747.

Phatak, P., Cookson, J., Dai, F., Smith, V., Gartenhaus, R., Stevens, M., and Burger, A. (2007). Telomere uncapping by the G-quadruplex ligand RHPS4 inhibits clonogenic tumour cell growth in vitro and in vivo consistent with a cancer stem cell targeting mechanism. Br. J. Cancer *96*, 1223-1233.

Porro, A., Feuerhahn, S., and Lingner, J. (2014). TERRA-reinforced association of LSD1 with MRE11 promotes processing of uncapped telomeres. Cell Rep. *6*, 765-776.

Portmann, S., Usman, N., and Egli, M. (1995). The crystal structure of r(CCCCGGGG) in two distinct lattices. Biochemistry *34*, 7569-7575.

Rajendran, A., Nakano, S.-i., and Sugimoto, N. (2010). Molecular crowding of the cosolutes induces an intramolecular i-motif structure of triplet repeat DNA oligomers at neutral pH. Chem. Commun. *46*, 1299-1301.

Redon, S., Reichenbach, P., and Lingner, J. (2010). The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. Nucleic Acids Res. *38*, 5797-5806.

Redon, S., Zemp, I., and Lingner, J. (2013). A three-state model for the regulation of telomerase by TERRA and hnRNPA1. Nucleic Acids Res. *41*, 9117-9128.

Riou, J., Guittat, L., Mailliet, P., Laoui, A., Renou, E., Petitgenet, O., Megnin-Chanet, F., Helene, C., and Mergny, J. (2002). Cell senescence and telomere shortening induced by a new series of specific G-quadruplex DNA ligands. Proc. Natl. Acad. Sci. U S A. *99*, 2672-2677.

Rodriguez, R., Muller, S., Yeoman, J.A., Trentesaux, C., Riou, J.F., and Balasubramanian, S. (2008). A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. J. Am. Chem. Soc. *130*, 15758-15759.

Salvati, E., Leonetti, C., Rizzo, A., Scarsella, M., Mottolese, M., Galati, R., Sperduti, I., Stevens, M.F., D'Incalci, M., and Blasco, M. (2007). Telomere damage induced by the G-quadruplex ligand RHPS4 has an antitumor effect. J. Clin. Invest. *117*, 3236-3247.

Schoeftner, S., and Blasco, M.A. (2007). Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat. Cell Biol. *10*, 228-236.

Schultze, P., Hud, N.V., Smith, F.W., and Feigon, J. (1999). The effect of sodium, potassium and ammonium ions on the conformation of the dimeric quadruplex formed by the *Oxytricha nova* telomere repeat oligonucleotide d(G4T4G4). Nucleic Acids Res. *27*, 3018-3028.

Simonsson, T. (2001). G-quadruplex DNA structures variations on a theme. Biol. Chem. *382*, 621-628.

Smargiasso, N., Rosu, F., Hsia, W., Colson, P., Baker, E.S., Bowers, M.T., De Pauw, E., and Gabelica, V. (2008). G-quadruplex DNA assemblies: loop length, cation identity, and multimer formation. J. Am. Chem. Soc. *130*, 10208-10216.

Sun, D., Thompson, B., Cathers, B.E., Salazar, M., Kerwin, S.M., Trent, J.O., Jenkins, T.C., Neidle, S., and Hurley, L.H. (1997). Inhibition of human telomerase by a G-quadruplex-interactive compound. J. Med. Chem. *40*, 2113-2116.

Völker, J., Klump, H.H., and Breslauer, K.J. (2007). The energetics of i-DNA tetraplex structures formed intermolecularly by d(TC5) and intramolecularly by d[(C5T3)3C5]. Biopolymers *86*, 136-147.

Vorlickova, M., Kejnovska, I., Bednarova, K., Renciuk, D., and Kypr, J. (2012). Circular dichroism spectroscopy of DNA: from duplexes to quadruplexes. Chirality *24*, 691-698.

Wang, G., and Vasquez, K.M. (2006). Non-B DNA structure-induced genetic instability. Mutat. Res. *598*, 103-119.

Warburton, P.E. (2004). Chromosomal dynamics of human neocentromere formation. Chromosome Res. *12*, 617-626.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids. Nature *171*, 737-738.

Wijmenga, S.S., and van Buuren, B.N. (1998). The use of NMR methods for conformational studies of nucleic acids. Prog. Nucl. Magn. Reson. Spectrosc. *32*, 287-387.

Xu, Y., Suzuki, Y., Ishizuka, T., Xiao, C.-D., Liu, X., Hayashi, T., and Komiyama, M. (2014). Finding a human telomere DNA–RNA hybrid G-quadruplex formed by human telomeric 6-mer RNA and 16-mer DNA using click chemistry: A protective structure for telomere end. Bioorg. Med. Chem. 22, 4419-4421.

Xu, Y., Suzuki, Y., Ito, K., and Komiyama, M. (2010). Telomeric repeat-containing RNA structure in living cells. Proc. Natl. Acad. Sci. U S A. *107*, 14579-14584.

Xu, Y., Suzuki, Y., and Komiyama, M. (2009). Click chemistry for the identification of G-quadruplex structures: discovery of a DNA-RNA G-quadruplex. Angew. Chem. Int. Ed. Engl. *48*, 3281-3284.

Yehezkel, S., Segev, Y., Viegas-Péquignot, E., Skorecki, K., and Selig, S. (2008). Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. Hum. Mol. Genet. *17*, 2776-2789.

Yoda, K., Kitagawa, K., Masumoto, H., Muro, Y., and Okazaki, T. (1992). A human centromere protein, CENP-B, has a DNA binding domain containing four potential alpha helices at the NH2 terminus, which is separable from dimerizing activity. J. Cell Biol. *119*, 1413-1427.

Zahler, A.M., Williamson, J.R., Cech, T.R., and Prescott, D.M. (1991). Inhibition of telomerase by G-quartet DMA structures. Nature *350*, 718-720.

Zhao, J., Bacolla, A., Wang, G., and Vasquez, K.M. (2010). Non-B DNA structure-induced genetic instability and evolution. Cell Mol. Life Sci. *67*, 43-62.

Zhou, J., Wei, C., Jia, G., Wang, X., Feng, Z., and Li, C. (2010). Formation of i-motif structure at neutral and slightly alkaline pH. Mol. BioSyst. *6*, 580-586.

Anexo

Otros artículos publicados durante el desarrollo de la tesis:

A minimal i-motif stabilized by minor groove G:T:G:T tetrads

Núria Escaja, Júlia Viladoms, Miguel Garavís, Alfredo Villasante, Enrique Pedroso y Carlos González

Nucl. Acids Res. (2012) 40 (22):11737-11747.doi: 10.1093/nar/gks911

Suplemmentary DATA FOR:

A minimal i-motif stabilized by minor groove G:T:G:T tetrads

Núria Escaja[2], Júlia Viladoms[2], Miguel Garavís[1,3], Alfredo Villasante[3], Enrique Pedroso[2,*] and Carlos González[1,*]

Tables and Figures mentioned in the main text:

Table S1. Assignment list of d<pTCGTTTCGTT> at pH 5.1, T=5ºC (25 mM phosphate buffer, 100 mM NaCl)
Table S2: Assignment list of d(TCGTTCGT) at pH 5, T=5ºC (25 mM phosphate buffer, 100 mM NaCl, 15 mM $MgCl_2$)
Table S3. Experimental constraints and calculation statistics for d<pTCGTTTCGTT>.

Figure S1. Schematic representation of G-tetrad (A), hemiprotonated $C:C^+$ base pair (B), and major and minor groove G:T:G:T tetrads (C and D, respectively).
Figure S2. NMR spectra of d<pTCGTTTCGTT> in H_2O/D_2O 9:1 at T = 5 ºC in 25 mM phosphate buffer, pH 7, 100mM NaCl. Top: low concentration (80 μM); Bottom: high concentration (800 μM).
Figure S3. Top: NMR spectra of d<pTCGTTTCGTT> at different temperatures and 80 μM oligonucleotide concentration. Bottom: NMR spectra of d(TCGTTTCGT) at different temperatures and 0.5 mM, 0.1 mM and 0.01 mM oligonucleotide concentration. All spectra are in H_2O/D_2O 9:1, 25 mM phosphate buffer, 100 mM NaCl, pH 5.0.
Figure S4: Non denaturing gel electrophoresis. (A) Non denaturing 20% PAGE in TBE buffer pH 8.3, 25 mM NaPi and 100 mM NaCl. (B) Non denaturing 20% PAGE in 10 mM Robinson-Britton buffer pH 4, 100 mM NaCl. Lanes: (**1**) dT ladders, (**2**) d(TCGTTTCGT), (**3**) d(AGCAAAGCA), (**4**) 1:1 mix of d(TCGTTTCGT) and d(AGCAAAGCA), (**5**) d(TCCGTTTCCGT), (**6**) TBA(Thrombin binding aptamer) d(GGTTGGTGTGGTTGG).
Figure S5. Exchangeable protons region of the NOESY spectrum (t_m=250 ms) of d(TCGTTTCGT) in H_2O/D_2O 9:1 in 25 mM phosphate buffer, pH 5, T=5ºC, 100 mM NaCl, 0.5 mM oligonucleotide concentration and schematic representations of proposed head-to-head and head-to-tail dimeric structures.
Figure S6. Exchangeable protons region of the NOESY spectrum (t_m=200 ms) of d(TCGTTCGT) in H_2O/D_2O 9:1 in 25 mM phosphate buffer, pH 5, T=5ºC, 100 mM NaCl, 10 mM $MgCl_2$, 0.66 mM oligonucleotide concentration, and a schematic representation of the proposed head-to-head dimeric structure**.**
Figure S7. Non-exchangeable protons region of the NOESY spectra of d(TCGTTCGT) (t_m=200 ms) (Left) and d(TCGTTTCGT) (t_m=250 ms) (Right) in H_2O/D_2O 9:1 in 25 mM phosphate buffer, pH 5, T=5ºC, 100 mM NaCl. Same oligonucleotide concentrations as in S5 and S6.
Figure S8. Proton connectivity map of d<pTCGTTTCGTT>.
Figure S9. Cytosine H6-H5 cross-peaks region of the TOCSY spectra of d(TCGTTTCGT) at different temperatures (25 mM phosphate buffer, pH 5, 100 mM NaCl, 0.5 oligonucleotide concentration).
Figure S10. NMR spectra of d<pTGCTTTGCTT> in H_2O/D_2O 9:1 in 25 mM phosphate buffer, 100 mM NaCl T= 5ºC. Top) pH 7.0 Bottom) pH 4.0, 0.5 mM oligonucleotide concentration.
Figure S11. Duplex competition experiments. NMR spectra of: (A) d(TCGTTTCGT) at pH 4.5, T=5ºC (100 μM oligonucleotide concentration, 25 mM phosphate buffer, 100 mM NaCl); (B) Complementary strand d(AGCAAAGCA) at pH 4.5, T=5ºC (100 μM oligonucleotide concentration, 25 mM phosphate buffer, 100 mM NaCl); (C, D and E) Equimolar mixture of d(TCGTTTCGT) and d(AGCAAAGCA), T=5ºC, 100 μM oligonucleotide concentration, 25 mM phosphate buffer, 100 mM NaCl at pH 7, 5 and 4.5, respectively.

SUPPLEMENTARY TABLES

Table S1. Assignment list of d<pTCGTTTCGTT> at pH 5.1, T=5ºC

Buffer conditions: 25 mM phosphate buffer, 100 mM NaCl

	H1/H3	H42/H22	H41/H21	H6/H8	H5/Me	H1'	H2'	H2"	H3'	H4'	H5'	H5"
T1	11.85	-	-	7.85	1.99	6.49	2.55		4.97	4.38	4.12	
C2	15.42	9.52	7.58	7.49	6.31	6.35	1.06	2.15	4.82	4.56	4.10	
G3	10.85	8.83	5.69	8.38	-	5.95	3.00	2.68	5.15	4.41	3.9, 4.11	
T4	11.5	-	-	7.63	1.76	6.08	2.07	2.33	4.81	3.99		
T5	10.47	-	-	7.77	1.92	6.43	2.28	2.55	4.63	4.47	3.98, 4.13	

Table S2. Assignment list of d(TCGTTCGT) at pH 5, T=5ºC

Buffer conditions: 25 mM phosphate buffer, 100 mM NaCl, 10 mM $MgCl_2$*

	H1/H3	H42/H22	H41/H21	H6/H8	H5/Me	H1'	H2'	H2"	H3'	H4'	H5'	H5"
T1	11.36	-	-	7.52	1.79	6.25	2.56	2.39	4.83	n.a.	n.a.	
C2	15.40	9.40	7.44	7.26	6.42	6.23	0.98	2.21	4.70	3.73	n.a.	
G3	10.36	8.64	6.00	8.14	-	5.80	2.76	2.51	4.95	n.a.	n.a.	
T4	**	-	-	7.34	1.51	5.72	1.85	2.02	4.27	n.a.	n.a.	
T5	**	-	-	7.61	1.74	6.11	2.40	2.52	n.a.	n.a.	n.a.	
C6	15.34	9.36	7.42	7.30	6.24	6.21	0.95	2.27	4.68	4.46	n.a.	
G7	10.10	n.o.		8.20	-	5.69	2.78	2.50	4.96	n.a.	n.a.	
T8	**	-	-	7.42	1.56	6.14	2.41, 1.97		n.a.	n.a.	n.a.	

*10 mM of $MgCl_2$ was added to the sample to obtain a higher thermal stability. Proton chemical shifts at T=5ºC do not differ from those obtained for the dimeric structure when only Na^+ was added.

**Chemical shifts of imino protons of T4, T5 and T8 are: 10.44, 10,71 and 11.21 ppm, but it was not possible to assign them.

n.o: not observed

n.a: not assigned

Table S3: Experimental constraints and calculation statistics of d<pTCGTTTCGTT>		
Experimental distance constraints		
Total number	234	
intra-residue	112	
sequential	70	
range > 1	52	
Intra-subunit	190	
Inter-subunit	44	
RMSD (Å)		
all well-defined bases[+]	1.1±0.3 Å	
all well-defined heavy atoms[+]	1.6±0.4 Å	
backbone	2.1±0.5 Å	
all heavy atoms	2.7±0.6 Å	
Residual violations	Average	Range
Sum of violation (Å)	3.4	2.4 .. 3.8
Max. violation (Å)	0.36	0.21 .. 0.47
NOE energy (kcal/mol)	12	8 .. 13
Total energy (kcal/mol)	-13500	-1219.. -2460

[+]All except unpaired thymines (4, 5, 9 and10).

SUPLEMENTARY FIGURES

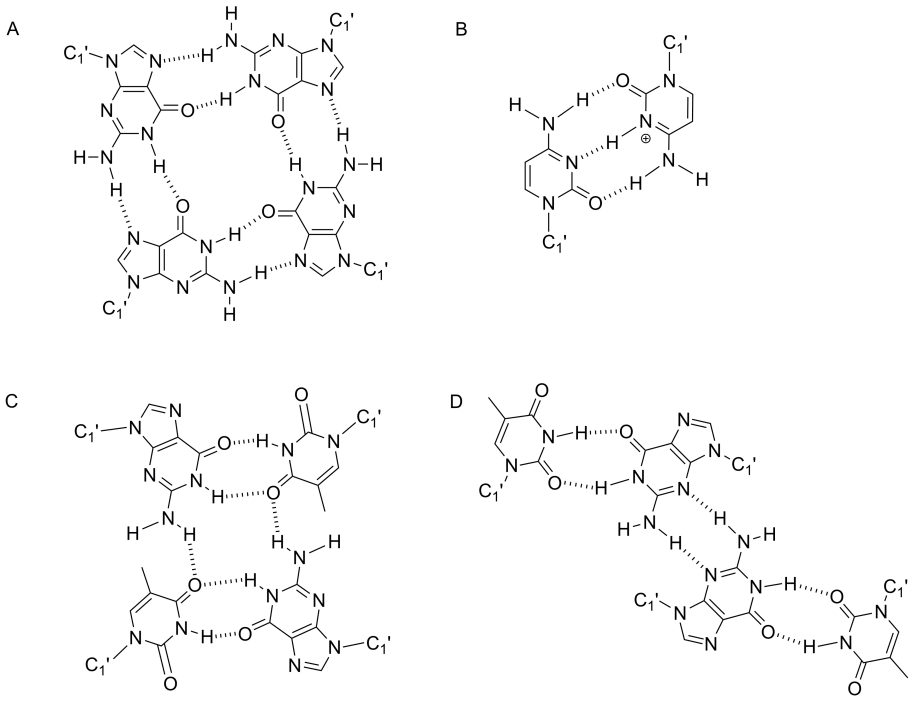


Figure S1. Schematic representation of G-tetrad (A), hemiprotonated C:C$^+$ base pair (B) and major and minor groove G:T:G:T tetrads (C and D, respectively).

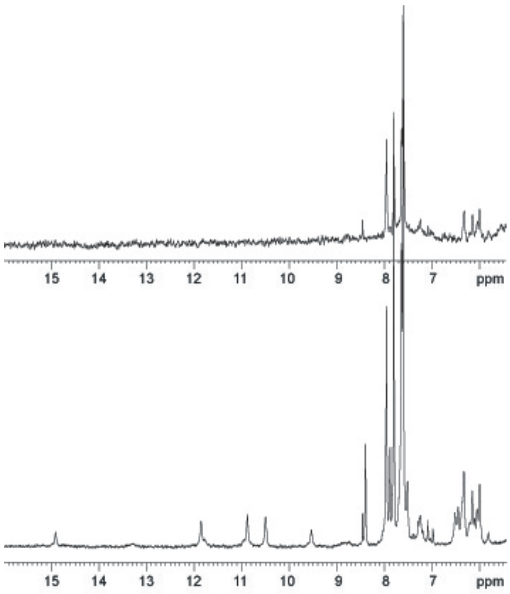


Figure S2. NMR spectra of d<pTCGTTTCGTT> in H_2O/D_2O 9:1 at T = 5 ºC in 25 mM phosphate buffer, pH 7, 100mM NaCl. Top: low concentration (80 μM); Bottom: high concentration (800 μM).

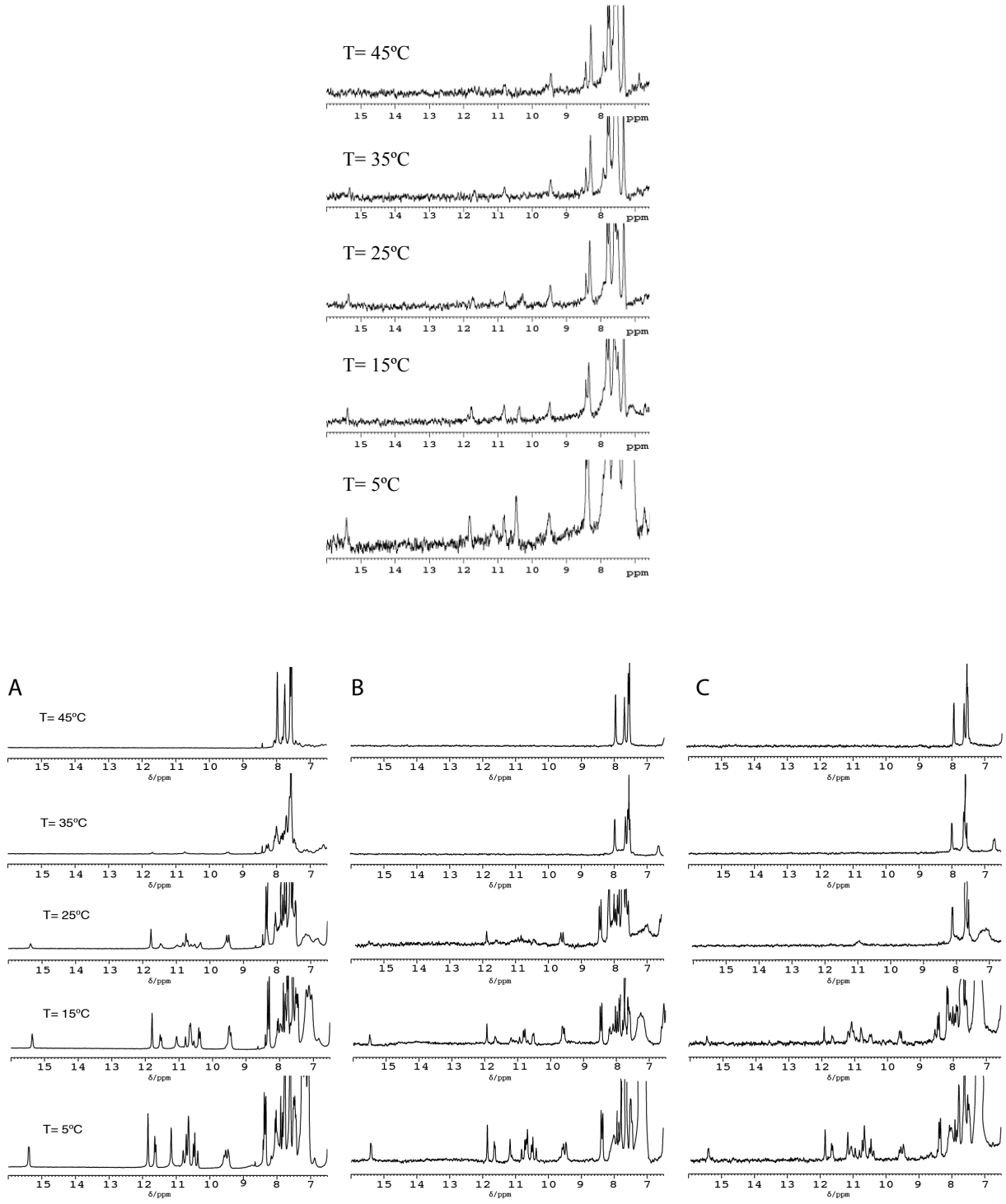


Figure S3. Top: NMR spectra of d<pTCGTTTCGTT> at different temperatures and 80 μM oligonucleotide concentration. Bottom: NMR spectra of d(TCGTTTCGT) at different temperatures and 0.5, 0.1 and 0.01 mM oligonucleotide concentration (A, B and C, respectively). All spectra are in H_2O/D_2O 9:1, 25 mM phosphate buffer, 100 mM NaCl, pH 5.0.

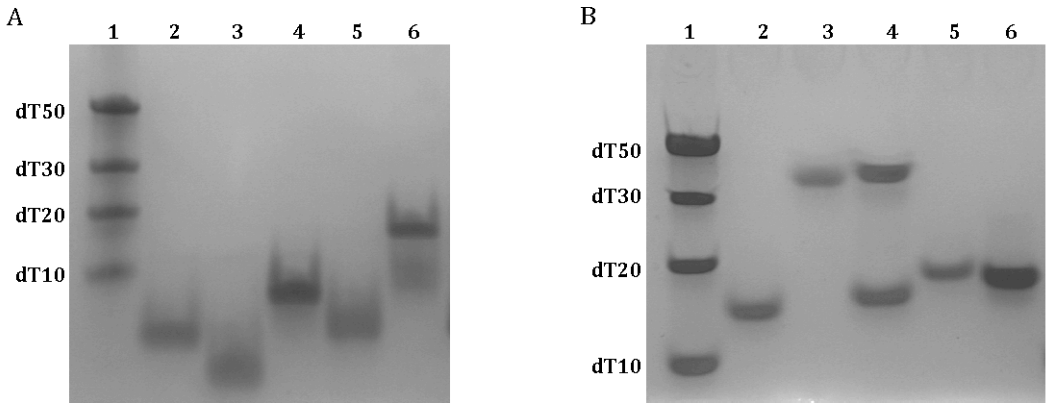


Figure S4. Non denaturing electrophoresis. (A) Non denaturing 20% PAGE in TBE buffer pH 8.3, 25 mM NaPi and 100 mM NaCl. (B) Non denaturing 20% PAGE in 10mM Robinson-Britton buffer pH 4, 100 mM NaCl. (**2**): d(TCGTTTCGT), (**3**): d(AGCAAAGCA), (**4**): 1:1 mix of d(TCGTTTCGT) and d(AGCAAAGCA), (**5**): d(TCCGTTTCCGT), (**5**): TBA (Thrombin binding aptamer) d(GGTTGGTGTGGTTGG). Lane (**1**): dT ladders.

Experimental methods on gel electrophoresis

Electrophoresis experiments were performed with 10 x 7cm native gel containing 20% polycrylamide (Acrylamide:Bis-acrylamide 19:1 ratio) in TBE Buffer pH 8.3 supplemented with 25 mM NaPi and 100 mM NaCl (Figure S4A) or in Robinson-Britton Buffer ($[CH_3COOH]= [H_3PO_4] = [H_3BO_3]$ =10 mM) pH 4 supplemented with 100 mM NaCl (Figure S4B). The samples were prepared at 80 μM concentration and were incubated overnight at 4ºC in TBE Buffer pH 8.3, 25 mM NaPi and 100 mM NaCl (Figure S4A) or in 40 mM Robinson-Britton Buffer pH 4 and 100 mM NaCl (Figure S4B). Gels were loaded after 1h preruning time and electrophoresis was performed at 10V/cm (Figure 4SA) or 8.5V/cm (Figure 4SB) for 90 and 180 min, respectively, at 4ºC. Gels were viewed after staining with Stains-All dye (Sigma E-9379).

Results

d(TCGTTTCGT) (lane 2) exhibits a larger mobility than TBA (a monomeric 15mer quadruplex) (lane 6) at pH 8.3, but their mobility is more similar at pH 4. This is consistent with monomeric and dimeric species of d(TCGTTTCGT) at pH 8.3 ant 4, respectively. d(TCCGTTTCCGT) (lane 5) is a variation of d(TCGTTTCGT) with two additional cytosines and exhibits a similar behaviour. This oligonucleotide most probably forms a dimeric i-motif with four hemiprotonated $C:C^+$, and it has been included in this experiment as an additional control.

d(AGCAAAGCA) (lane 3) presents very different mobility at different pHs. At neutral pH is monomeric and most probably unstructured, but it forms some higher order structure at acidic pH (maybe tetrameric).

The equimolar mix of d(TCGTTTCGT) and d(AGCAAAGCA) (lane 4) exhibits a single spot at pH 8 (a single duplex structure). However, two species are clearly observed at pH 4, consistent with the dimeric i-motif described in this manuscript, and a higher order structure adopted by d(AGCAAAGCA).

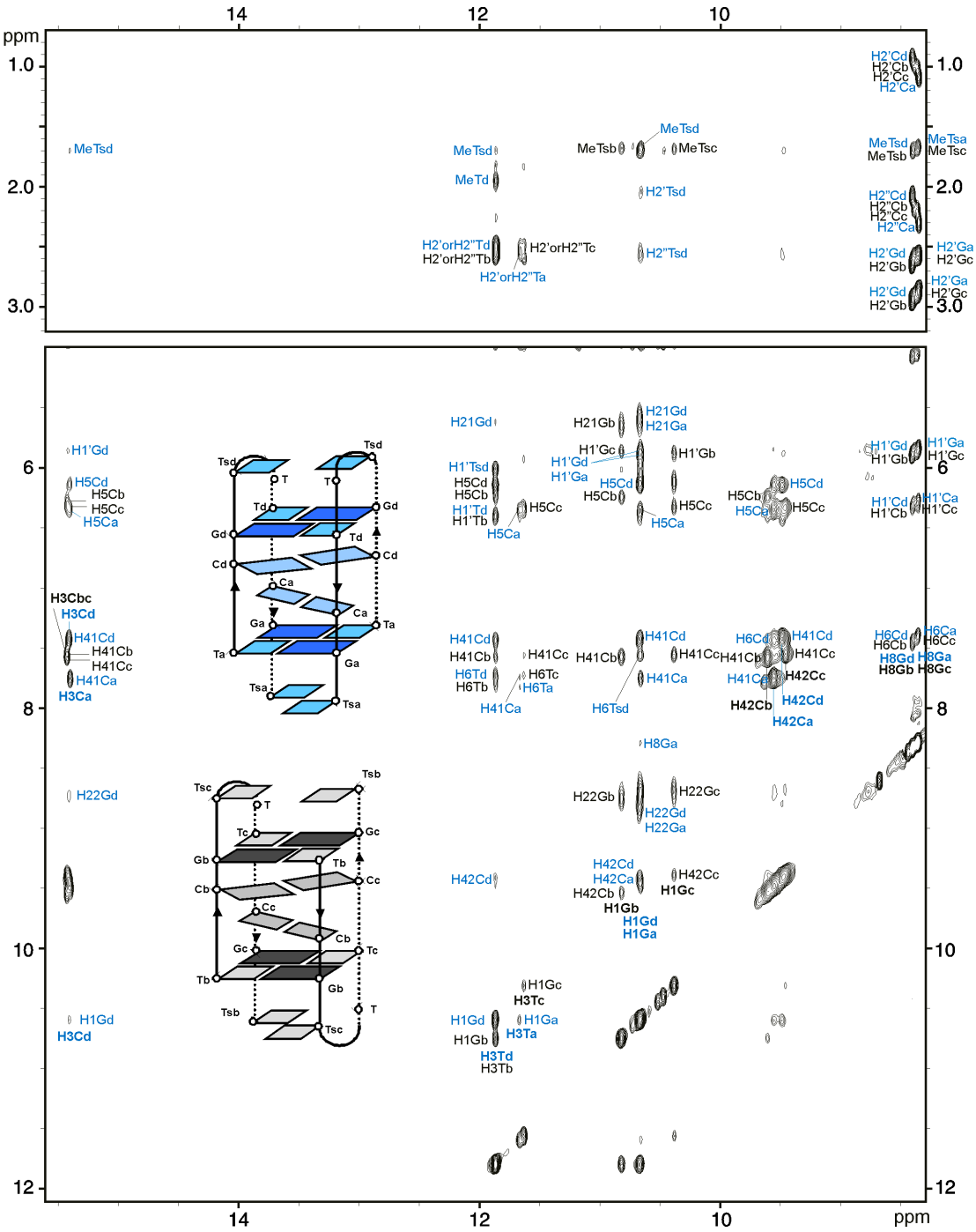


Figure S5. Exchangeable protons region of the NOESY spectrum (t_m = 300 ms) of d(TCGTTTCGT) in H_2O/D_2O 9:1 in 25 mM phosphate buffer, pH 5, T=5ºC, 100 mM NaCl, 0.5 mM oligonucleotide concentration and schematic representations of proposed head-to-head and head-to-tail dimeric structures. Signals corresponding to the head to head orientation are labelled in blue and the signals corresponding to the head to tail dimer are labeled in black.

Thymine residues are highly overlapped as only six Me-H6 cross-peaks can be distinguished in the TOCSY spectra (not shown). Only those thymine residues that show stacking interactions or are involved in the G:T:G:T tetrads can be identified and labeled according to their contacts with guanine and cytosine residues. Four spin systems corresponding to cytosines are clearly observed, both in the exchangeable and non exchangeable proton region (Figures S7 and S9). These residues exhibit imino-amino proton cross-peaks in the characteristic region of hemiprotonated $C:C^+$ base pairs (labeled Ca,Cb,Cc and Cd in the Figure). Two of these $C:C^+$ base pairs are formed by equivalent cytosines (Ca and Cd, head to head orientation) and one of them is formed by two different cytosines (Cb and Cc, head to tail orientation). Four spin systems correspond to guanines and show cross-peaks that reveal their implication in minor groove G:T:G:T tetrads (H1-H3 imino-imino proton cross-peaks between guanine and thymine residues and guanine amino-H1' contacts through the minor groove). Based on guanine amino-H1' cross-peaks, two of the G:T:G:T tetrads are formed by two equivalent guanines (head to head orientation) and one of the tetrads is formed by two non equivalent guanines (head to tail orientation).

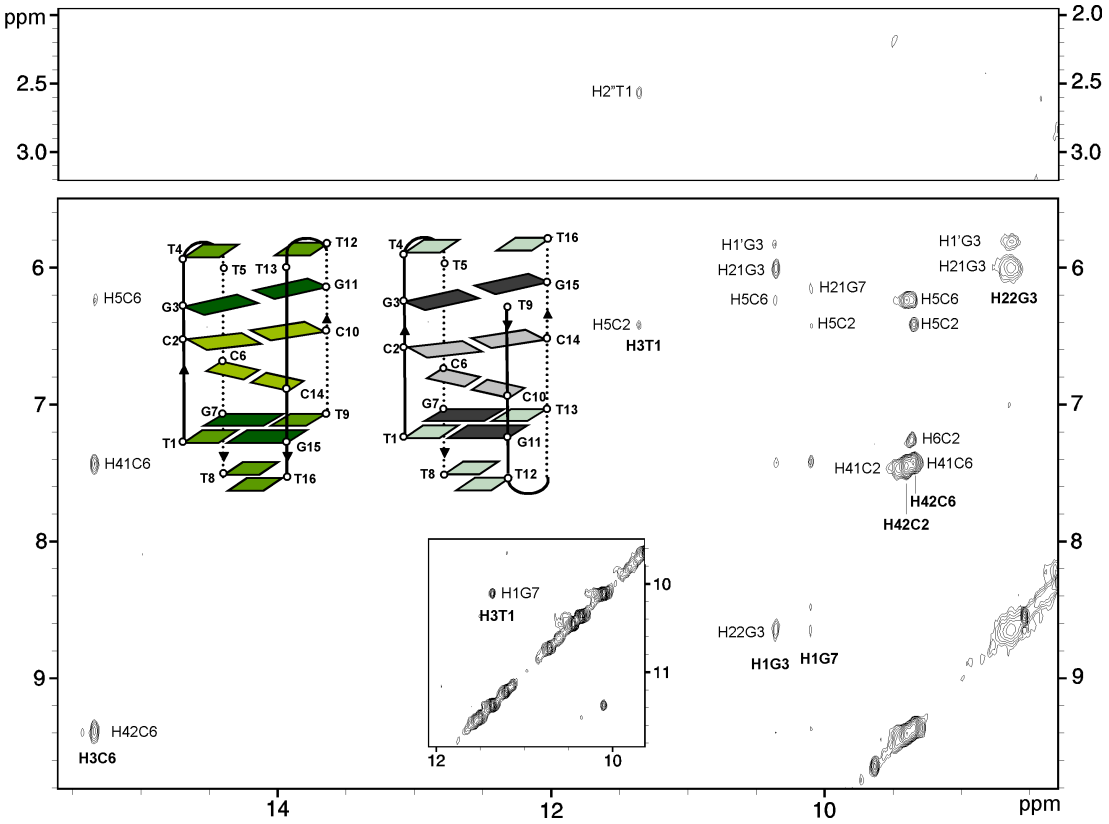


Figure S6. Exchangeable protons region of the NOESY spectrum (t_m = 200 ms) of d(TCGTTCGT) in H_2O/D_2O 9:1 in 25 mM phosphate buffer, pH 5, T=5ºC, 100 mM NaCl, 15 mM $MgCl_2$, 0.66 mM oligonucleotide concentration, and schematic representation of the proposed head-to-head dimeric structure (green coloured) and the alternative not observed head-to-tail structure. According to the observed H3-H41/H42/H5 crosspeaks, the most stable hemiprotonated C:C^+ base pair is the one formed between equivalent cytosine residues. This rules out the head-to-tail orientation, since the H3 proton of protonated cytosine would show cross-peaks with the amino protons of a non equivalent base-paired cytosine (C6 and C10).

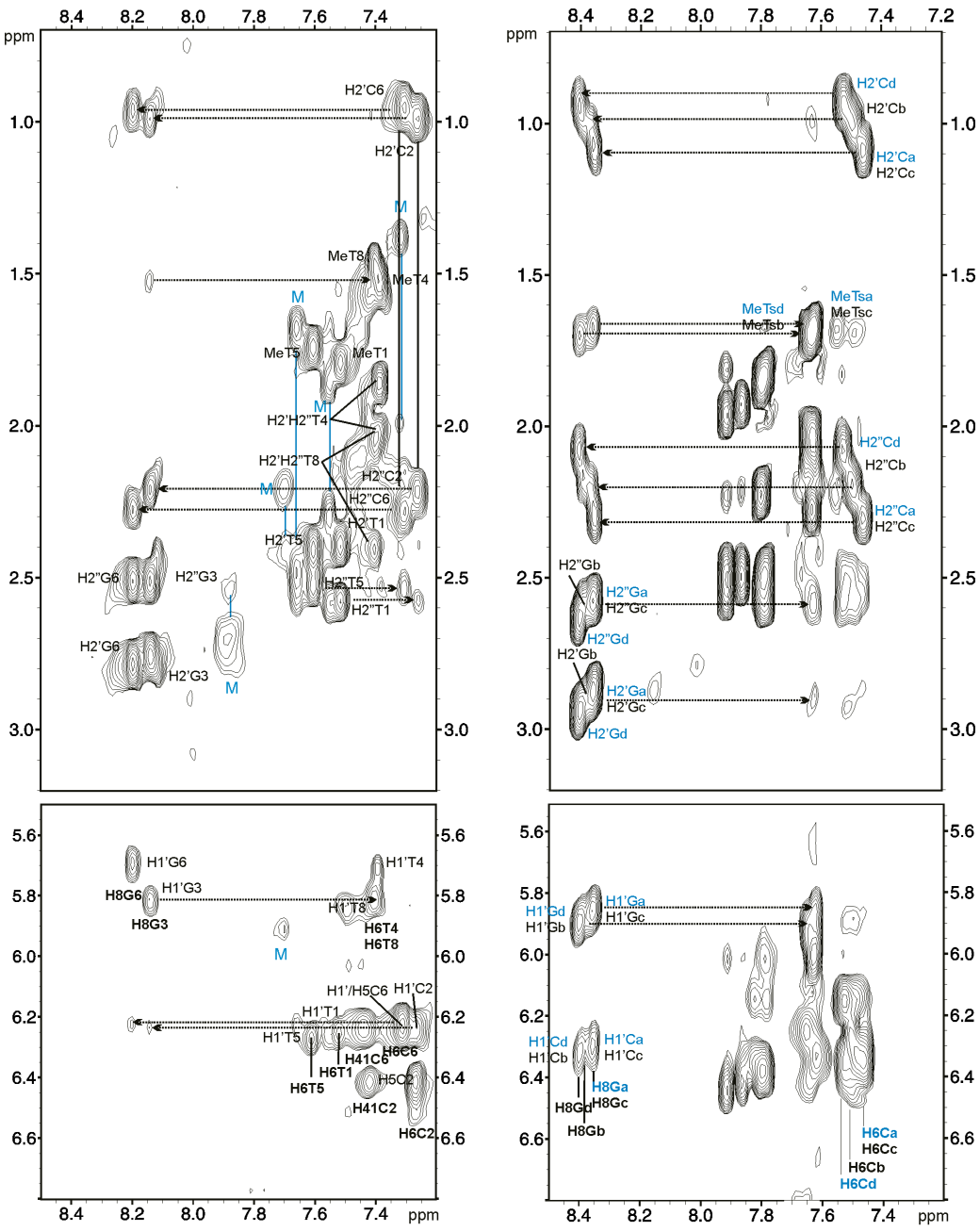


Figure S7. Non-exchangeable protons region of the NOESY spectra of d(TCGTTCGT) (t_m = 200 ms) (Left) and d(TCGTTTCGT) (t_m = 250 ms) (Right) in H_2O/D_2O 9:1 in 25 mM phosphate buffer, pH 5, T=5ºC, 100 mM NaCl. Signals corresponding to the unstructured species of d(TCGTTCGT) have been labelled with an M (left panel). In the right panel, signals corresponding to the head-to-head and the head-to-tail dimeric structures of d(TCGTTTCGT) are labelled in blue and black, respectively. Oligonucleotide concentrations are the same as in S5 and S6.

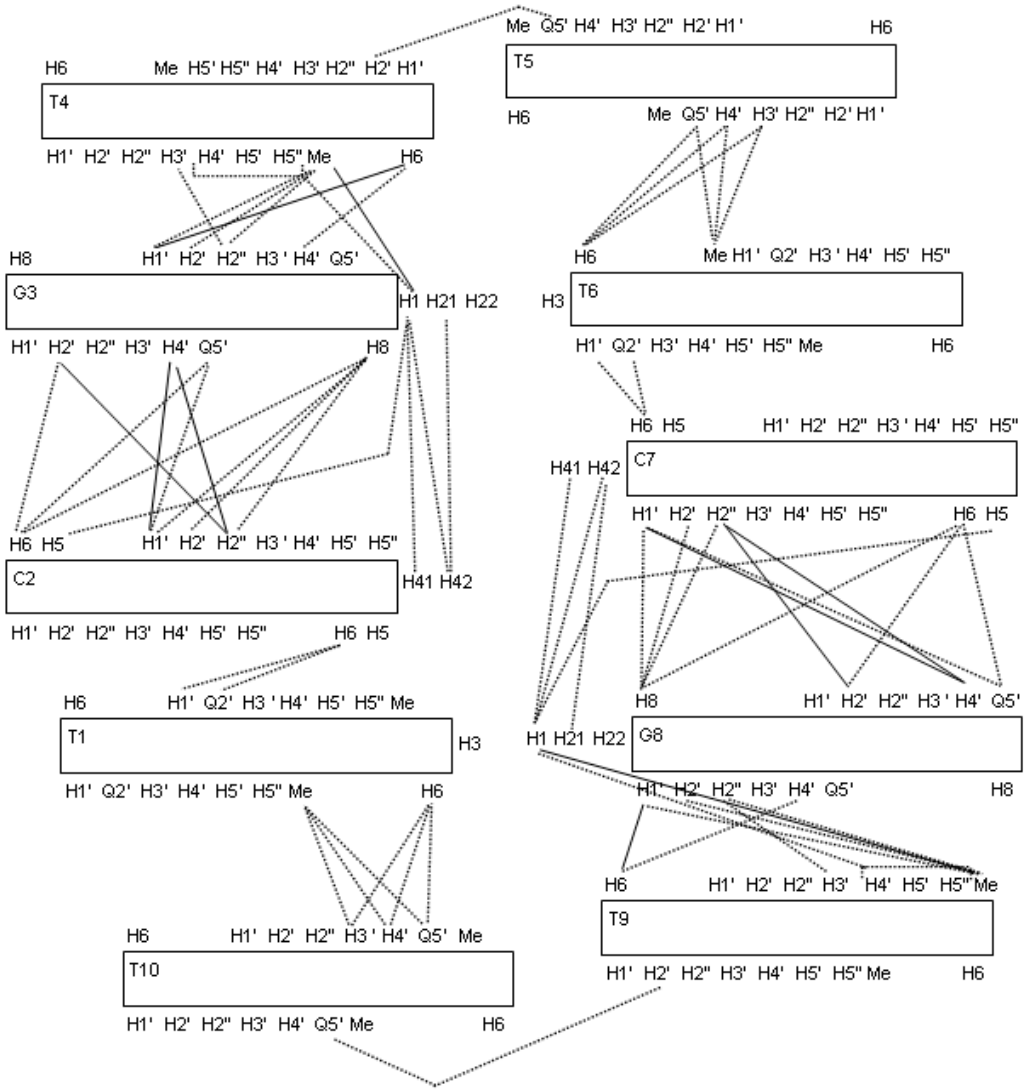


Figure S8. Proton connectivity map of d<pTCGTTTCGTT>.

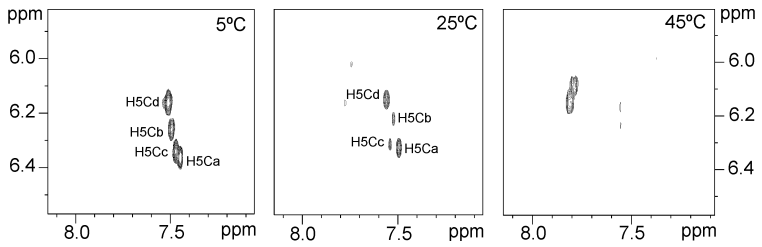


Figure S9. Cytidine H6-H5 cross-peaks region of the TOCSY spectra of d(TCGTTTCGT) at different temperatures (25 mM phosphate buffer, pH 5, 100 mM NaCl, 0.5 oligonucleotide concentration).

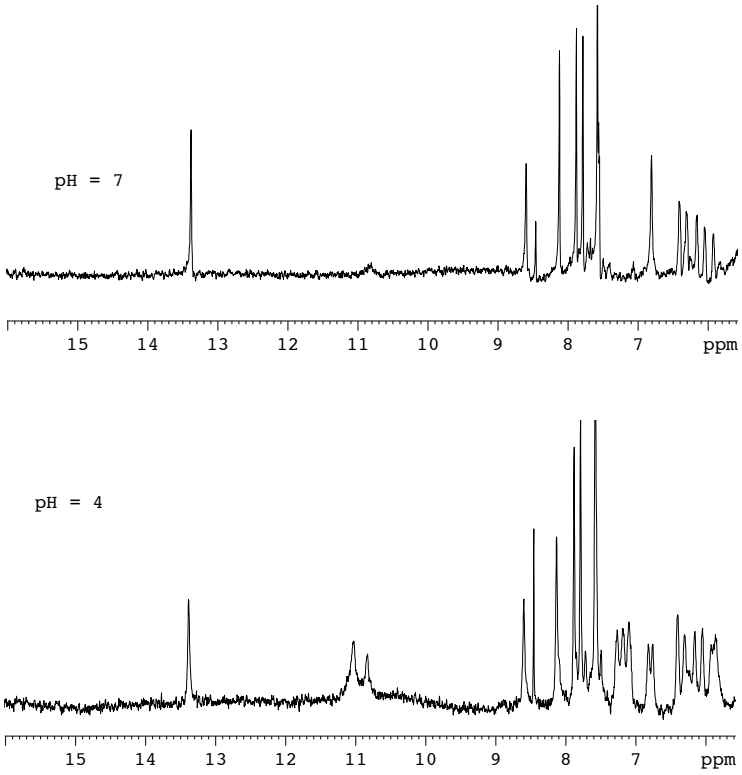


Figure S10. NMR spectra of d<pTGCTTTGCTT> in H_2O/D_2O 9:1 in 25 mM phosphate buffer, 100 mM NaCl T= 5ºC. Top) pH 7.0 Bottom) pH 4.0. 0.5 mM oligonucleotide concentration.

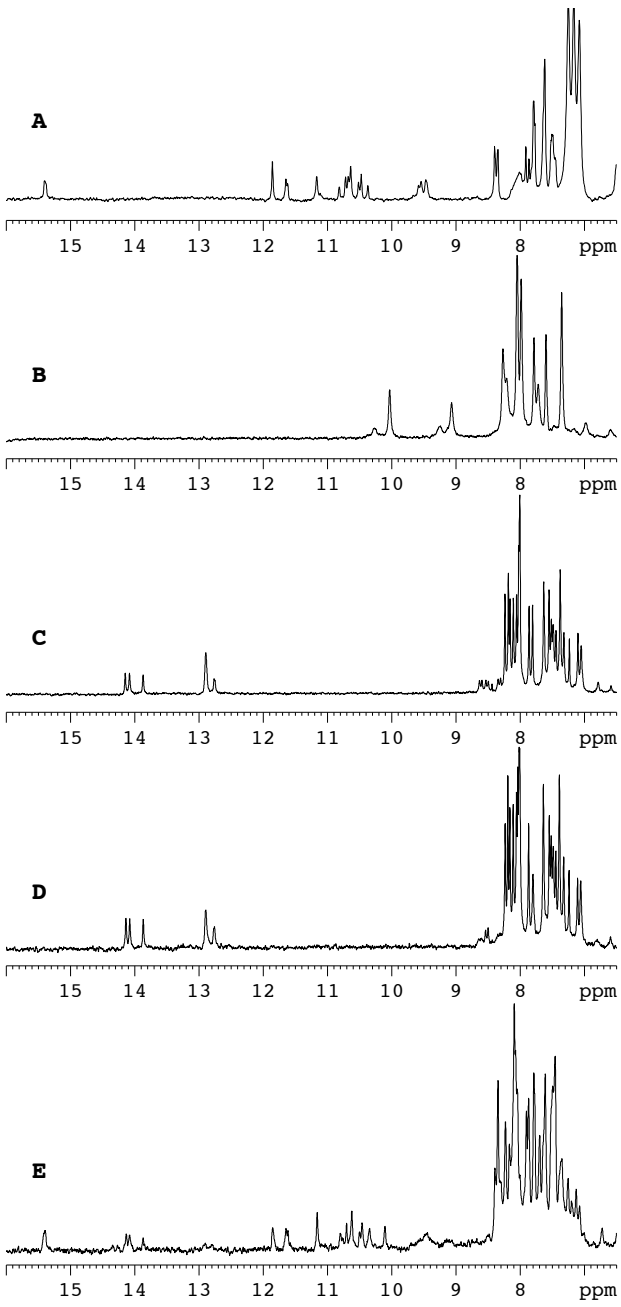
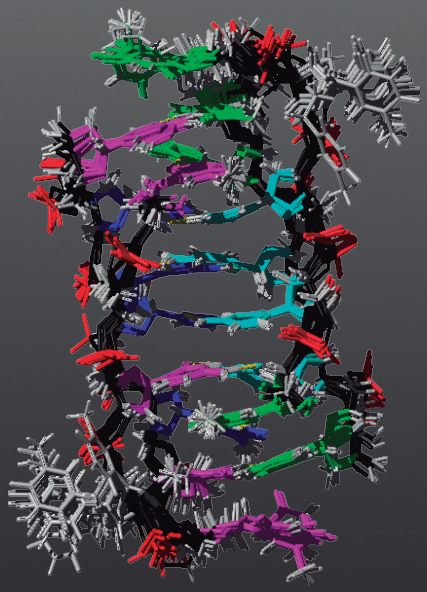


Figure S11. Duplex competition experiments. NMR spectra of: (A) d(TCGTTTCGT) at pH 4.5, T=5ºC (100 μM oligonucleotide concentration, 25 mM phosphate buffer, 100 mM NaCl); (B) Complementary strand d(AGCAAAGCA) at pH 4.5, T=5ºC (100 μM oligonucleotide concentration , 25 mM phosphate buffer, 100 mM NaCl); (C, D and E) Equimolar mixture of d(TCGTTTCGT) and d(AGCAAAGCA), T=5ºC, 100 μM oligonucleotide concentration, 25 mM phosphate buffer, 100 mM NaCl at pH 7, 5 and 4.5, respectively.



Instituto de Química Física Rocasolano
Centro de Biología Molecular Severo Ochoa
CSIC
Madrid, febrero 2015